MULTIMODAL STORY COMPREHENSION: DATASETS, TASKS AND NEURAL MODELS

By

HAREESH RAVI

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computer Science

Written under the direction of

Mubbasir Kapadia

And approved by

____________________________________

____________________________________

____________________________________

____________________________________

New Brunswick, New Jersey

October 2021
ABSTRACT OF THE DISSERTATION

Multimodal Story Comprehension: Datasets, Tasks and Neural Models

by Hareesh Ravi

Dissertation Director: Prof. Mubbasir Kapadia

Storytelling is a uniquely human skill that plays a central role in how we learn about and experience the world. Stories play a crucial part in the mental development of humans as it helps us encode a wide range of shared knowledge, including common sense physics, cause and effect, human psychology, and morality. We postulate that machine intelligence requires comparable skills, particularly when interacting with people. Much of the current research in understanding the visual and textual world operates only at a superficial, factual level using data that align atomic one–to–one descriptive, factual text with an image. An ideal AI system must be able to create and comprehend multimodal narratives in a causal and coherent manner, much like humans, to be able to seamlessly interact with humans. This dissertation aims to bridge the gap between current research and ideal AI systems by developing novel datasets, tasks and neural methods for true multimodal story creation and comprehension.

We start by highlighting limitations of existing works such as factual and superficial alignment of image–text context, lack of coherent narrative understanding and ill–defined tasks for multimodal story comprehension. To offset these limitations, we propose a novel computational task, Story Illustration, as a measure of story comprehension by a neural model. We model textual coherence explicitly with an end–to–end trained hierarchical neural model capable of illustrating a story. Our evaluation highlights limitations of ex-
isting visual storytelling datasets. Then, we extend the formulation to a Many–to–Many setting to generalize the task and demand coherence modelling by also creating a new and improved dataset for multimodal story comprehension. We develop a machine translation approach to story illustration leveraging text generation techniques that explicitly models visual and textual coherence in stories. Further, we develop a novel dataset AESOP that enables the modelling of story comprehension and creation from a truly multimodal perspective, capturing the creative process associated with storytelling. Our framework models the evolution of textual and visual concepts on AESOP using interacting sequential networks.

Our contributions lay a strong foundation for developing AI systems that can create and comprehend multimodal stories and consequently comprehend any kind of data. This dissertation along with all the publicly released resources such as data and code drives further research towards building complex and intelligent systems.
I would like to take this opportunity to express my gratitude to some important people in my life. This dissertation and my PhD would not be possible without the help, support, guidance and love I have received from them.

First and foremost, I am incredibly lucky to have had Dr. Mubbasir Kapadia as my advisor. He was crucial in getting me admission into the PhD in CS program at Rutgers. Since the beginning, he has been a guiding light throughout this journey with his insightful and honest reviews providing a perfect mix of active guidance with weekly meetings and absolute freedom to pursue my research ideas. I will forever be grateful to have worked with him.

I am grateful to Dr. Gerard De Melo for his valuable advising. A part of this dissertation has been completed under his active guidance. He helped me immensely through the toughest two years of my PhD and advised me on numerous occasions providing ideas on various aspects of research, life and career.

I thank my committee members Dr. Dimitris Metaxas and Dr. Nasrin Mostafazadeh for providing valuable suggestions and reviews about my dissertation and research. I owe special thanks to Dr. Leonid Sigal for his wonderful guidance, insightful research discussions and for inspiring me to be a good human being and researcher. I also thank Dr. Amelie Marian for being on my qualifying exam committee and Dr. Dimitris Metaxas, Dr. Ashutosh Modi, Dr. Scott Cohen, Dr. Kushal Kafle, Dr. Jonathan Brandt, Dr. Matthew Stone, Dr. Vladimir Pavlovic and Dr. Linda Ness for their amazing research collaboration. I am thankful to my colleagues Honglu, Malihe, Rajarshi and Sreyasi for their collaboration and the numerous discussions that kept me excited about research. I am also grateful to the extremely supportive staff, Carol, Michelle, Frederick and Robert, the International Student Office and their advisors and faculty at Rutgers who have helped me through this journey.
I express my sincere gratitude to Dr. A. V. Subramanayam for being my advisor during my time in IIIT-Delhi. His support and guidance were the foundation to my PhD. I also thank Dr. Ponnurangam Kumaraguru and my colleagues at IIIT-Delhi, Venkatesh, Archit, Tejas Bhai, Samarth, Anush, Megha and Ajay for the week of paradoxes, PCA simplification, beat box sessions and numerous intellectual discussions about god, the universe and everything. Special thanks to my wonderful friend Anupama for forcefully getting me out of my comfort zone, helping me through my career and all the amazing food that we explored together.

I thank my close friend and research companion, Sabarish (Oh!) for being a bouncing wall of ideas and a constant support for the past 10 years, both academically and personally. I thank my dear friends Akshaya, Divya, Gopinath, Guru, Rajesh, Sai and Shiva for being an important source of happiness throughout my life since kindergarden.

I would like to thank my lovely wife Srividya for making my day-to-day life worth living. She showers me with constant love and support and ensures my health and well-being despite pursuing her own PhD.

I am also incredibly grateful to my parents Brindha and Ravi for instilling the value of education in me since early childhood. Even through extreme personal and financial hardships, they did not fail to give me the best in education. They believed me enough to let me choose and do things that they did not fully comprehend. I am lucky and honored to be their son. I can not express in words my gratitude to my anna Prakash and manni Sowmya. They have made many personal sacrifices, supported and been with me throughout my life and never at any point lost their trust and belief in me. They are the major reason I am at this juncture today. Last but not the least, thank you Nilan, our little Gulab Jamun, for having born and giving us immense joy.
To my family. Thank you for everything.
# TABLE OF CONTENTS

Abstract ................................................................. ii

Acknowledgments ......................................................... iv

Dedication ................................................................. vi

List of Tables ............................................................. xiii

List of Figures ........................................................... xvi

Chapter 1: Introduction ................................................... 1

1.1 Why Study Stories? ................................................... 2

1.2 Limitations of Prior Work .......................................... 2

1.3 Proposed Approach .................................................. 3

1.4 Summary of Contributions ........................................... 4

1.5 List of Relevant Publications ....................................... 5

Chapter 2: Background and Prior Work ................................. 6

2.1 Language Generation/Retrieval For Visual Input .................. 7

2.1.1 Image to Text ..................................................... 8

2.1.2 Video to Text ..................................................... 8

2.1.3 Beyond Descriptive Captions .................................... 9
4.4 Experimental Setup ........................................... 29
  4.4.1 Training Details ........................................ 29
  4.4.2 Comparative Evaluation ................................. 31
  4.4.3 Visual Saliency Based Metric ......................... 32
  4.4.4 User Study ............................................. 34

4.5 Results and Discussion ..................................... 35
  4.5.1 User Study ............................................. 35
  4.5.2 Importance of Consistency and Coherence ........... 38
  4.5.3 Visual Saliency Metric ................................. 39

4.6 Limitations and Improvements ......................... 40

Chapter 5: Modelling Coherence and Many-to-Many Mapping .... 41
  5.1 Generalizing Story Illustration to Many-to-Many .... 42
  5.2 MMSI Dataset ............................................ 46
    5.2.1 Necessity ......................................... 46
    5.2.2 Description ....................................... 46
    5.2.3 Dataset Evaluation ................................. 49
  5.3 Machine Translation for MMSI ............................ 50
    5.3.1 Text Encoder ..................................... 51
    5.3.2 Image Encoder ................................... 52
    5.3.3 Text-to-Image Decoder ......................... 52
    5.3.4 Loss Function ................................... 53
    5.3.5 Sequential Image Retrieval Techniques .... 53
5.3.6 Training Methodologies ........................................ 54
5.4 Experimental Setup .............................................. 55
  5.4.1 Training Details ........................................... 55
  5.4.2 Evaluation Metrics ....................................... 56
  5.4.3 Baselines and Comparisons ............................... 56
5.5 Results and Discussion ........................................ 58
  5.5.1 MMSI .................................................. 58
  5.5.2 MMSC .................................................. 58
  5.5.3 User Study ............................................. 58
  5.5.4 Qualitative Analysis .................................... 64
  5.5.5 Training and Decoding Techniques ....................... 65
  5.5.6 Generalization to VIST ................................ 65
5.6 Limitations and Improvements ................................ 66

Chapter 6: Multimodal Creative Composition .................. 68
  6.1 Redefining Visual Storytelling with AESOP ............... 69
  6.2 AESOP Dataset ........................................... 71
    6.2.1 Guiding Principles .................................. 71
    6.2.2 Data Acquisition ..................................... 72
    6.2.3 Data Acquisition Web Interface ....................... 77
    6.2.4 Data Statistics ...................................... 78
    6.2.5 Example Stories ..................................... 81
    6.2.6 AESOP Vs. Other datasets ............................ 81
6.3 Towards Comprehension and Creative Composition of visual stories with AESOP ........................................... 87
  6.3.1 Assistant Illustrator ............................................... 88
  6.3.2 Assistant Writer ................................................... 88
6.4 AESOP The Model .................................................... 89
  6.4.1 Abstract Visual Representation ................................. 89
  6.4.2 Story Encoder ..................................................... 91
  6.4.3 Panel Decoder ..................................................... 93
6.5 Experimental Setup .................................................. 95
  6.5.1 Baselines and Comparison ........................................ 95
  6.5.2 Evaluation ........................................................ 96
  6.5.3 Training Details ................................................... 97
6.6 Results ............................................................... 98
  6.6.1 Assistant Illustrator ............................................... 98
  6.6.2 Assistant Writer .................................................. 104
  6.6.3 Model Limitations and Future Work ......................... 108
  6.6.4 Inadequacies of Automatic Metrics .......................... 108
  6.6.5 Complexity of AESOP ........................................... 108
6.7 Limitations and Improvements ..................................... 109

Chapter 7: Image–Text Coherence ..................................... 110
  7.1 Exploiting Cross–Modal Coherence for Text–to–Image Retrieval ................................. 110
  7.2 Datasets ................................................................ 114
  7.3 Cross Modal Coherence Model ................................. 115
LIST OF TABLES

4.1 The results of pairwise preference test on story visualization of workers reviews via AMT. Comparisons are conducted in the manner of A vs. B. The numbers indicates the percentage of responses that A is a better visualization than B for a given story. 37

4.2 Preference of algorithm based on maximum voting of 5 workers for 200 samples. To avoid ties, if GT and CNSI get 2 votes each and BL gets one, then both GT and CNSI gets half a point. 37

4.3 Visual Saliency based Recall@1, 2 and 5. 39

4.4 Visual Saliency based Recall of GT images@10, 50, 100 and 500. 39

5.1 Description of the models used for comparison. The prefix ‘P’ indicates that the proposed model is used except for the indicated change. In ‘Baseline’ model, Decoder attention is removed and Hierarchical Sentence Story Encoder is replaced with a Bi-GRU. P_Teacher model is trained with Teacher Forcing. 57

5.2 Model evaluation on Story-Illustration and Story-Completion (indicated by the suffix t_1 meaning first image is given as seed during inference and the model predicts the remaining) tasks. Med-Rank is calculated over the entire 35 164 images and the number in brackets indicate Med-Rank with a retrieval range [10] of 1000. 59

5.3 Baseline and Proposed models evaluated on Story-Completion task. t_i indicates i^{th} ground truth image is provided. Only subset of metrics are provided owing to space constraints. 65

5.4 Models trained and tested on VIST dataset. t_1 and t_{1,2,3} indicate story completion experiments. 66

6.1 Object and Word statistics of AESOP dataset. 78
6.2 Comparison with other datasets. Verb Frequency is the percentage of verbs over all words in the text. Top 30 verbs is the percentage of top 30 verbs over all verbs. Visible and Non-visible verbs indicate the frequency of select words per million words.

6.3 Results of all models on Assistant Illustrator and Assistant Writer modes when the TOP: last panel is masked and BOTTOM: middle panel is masked. For Assistant Illustrator, we provide accuracy over entire test set for prediction of BG (background) Dep (z value), Flip, Pose and Expr (Expression). Loc is the location similarity while O-IOU is the intersection over union between predicted and ground truth set of objects. Metrics for object attributes are calculated only if the predicted object is present in ground truth. Scene is the scene similarity metric. For Assistant Writer mode, B–1 indicates BLEU–1, B–4 is BLEU–4, M is METEOR, R–L is ROUGE–L and C is CIDEr.

6.4 Results of user study comparing models pairwise along three dimensions for Assistant Illustrator. Values are given in % and overall indicates the overall preference between the two shown models.

7.1 Coherence relations, their distribution and entropy in CITE++ dataset. We use the question identifier and the relation name interchangeably. Positive rate is the percentage of samples that are labeled as ‘Yes’ for that question.

7.2 Coherence relations and their distribution in Clue dataset [13].

7.3 Description of the models used for comparison. -NoAttn means removing the attention module from the proposed model. ‘All’ relations indicate that the Coherence Aware Module is trained with all the relations in a multi-label multi-class setting. c indicates only one relation is used with the Coherence Aware Module in a binary classification setting.

7.4 Quantitative comparison of the models trained and evaluated on Clue dataset.

7.5 Quantitative comparison in CITE++ dataset. The relations corresponding to each $Q_i$ are shown in Table 7.1. ↓ indicates that lower the better and ↑ indicates that higher the better.

7.6 Human evaluation results. Values indicate the percentage of samples for which humans preferred the output of CMCM, CMCA, both or neither.

7.7 Average Precision of Coherence relation prediction using probabilities from the Coherence Aware Module.
7.8 Average Precision of Coherence relation prediction using probabilities from the *Coherence Aware Module*. All models are the proposed CMCM variants. For relation specific models, the prefix CMCM is removed due to space constraints.
LIST OF FIGURES

3.1 Illustration showing the story illustration task. Input is surrounded by a red box while output is in green box. The blue arrows indicate the reasoning pathways for the proposed model and tasks. ........................................... 16

3.2 Illustration showing the Many–to–Many story illustration task. Input is surrounded by a red box while output is in green box. The blue arrows indicate the reasoning pathways for the proposed model and tasks. ............... 17

3.3 Illustration showing the AESOP Assistant tasks. Input is surrounded by a red box while output is in green box. The blue arrows indicate the reasoning pathways for the proposed model and tasks. .......................... 18

4.1 Two image sequences visualize the given story (top). The images predicted by the proposed sequential model with coherence (bottom) demonstrate higher consistency and better alignment with the story than the images retrieved independently sentence-by-sentence (middle). ......................... 21

4.2 Two samples of ground truth (middle) and our prediction (bottom) for the story (top). The bottom sequence of images win more votes from AMT workers for better visual illustration. ........................................... 24

4.3 The proposed approach for story visualization: (a) Shows modeling of isolated sentences for sentence encoding and parse tree extraction for coherence vector computation. (b) Uses encoded sentence vectors to train sentence-RNN using Order Embedding (OE) loss function [101]. (c) Uses encoded sentence and image vectors along with coherence vector to sequentially encode the input story. Story-RNN is trained using modified OE-loss function with On-line Negative Mining(ONM) for generating negative samples within a batch. ........................................... 26

4.4 Two image sequences visualize a ”graduation” story. AMT workers prefer the GT over BL, though both look similar. ........................................... 31
4.5 A sample that our proposed method (bottom) gets less votes from AMT users than the ground truth (middle). However the two image sequences look visually similar. ................................. 33

4.6 Example story where CNSI failed to better illustrate the story than BL according to the user study. ................................. 35

4.7 Samples of predicted images for three stories where the image sequences in the last row predicted by our CNSI model wins the most votes from AMT workers. ................................. 36

4.8 The Amazon workers prefer the predicted image sequence in the bottom which shows less visual consistency than the middle one. ................................. 38

5.1 Example input story and corresponding illustration (ordered row wise starting from top left) from the created dataset. Note the coherence between images in the illustration. ................................. 43

5.2 An Example story from the proposed dataset. High visual coherence and visual grounding is visible. The text is detailed yet narrative in nature capturing the storytelling intent. ................................. 45

5.3 More example stories from the proposed MMSI dataset. All the stories show high visual and textual coherence. ................................. 48

5.4 An example story with intricate details (2nd and 3rd sentences) in the text not explicitly captured in the visual. ................................. 49

5.5 Proposed Encoder-Decoder training framework for story illustration. (A) is the sentence encoder (B) story encoder (C) text-to-image decoder and (D) is the trainable part of image encoder. During inference, image closest to $\hat{I}_t$ (generated by text-to-image decoder) is retrieved and then given as input for retrieving subsequent images. ................................. 51

5.6 An example input story, ground truth and the predicted sequences of images by the proposed model with $t_1$ followed by $t_{1,10,15}$. BLUE border indicates seed images, GREEN border indicates exact match to ground truth while YELLOW border indicates the presence of the predicted image somewhere in the ground truth story. No border means the predicted image is not part of the ground truth. BLACK image is the $\langle EOS \rangle$. Recall@1 for $t_1$ is calculated as $\# \text{GREEN} / (\# \text{GT} - \# \text{BLUE}) = 37.5\%$. Similarly StoryRecall@1 is calculated as $(\# \text{GREEN} + \# \text{YELLOW}) / (\# \text{GT} - \# \text{BLUE}) = 50.0\%$. The Recall@1 and StoryRecall@1 for $t_{1,10,15}$ are 31.8\% and 54.5\% respectively. MAE for this example is 0.0 for both the predictions. ................................. 61
5.7 Another example result. The Recall@1 and StoryRecall@1 for $t_1$ and are 54.5% and 100% while for $t_{1,10,15}$ are 54.5% and 95.4% respectively. MAE for this example is 0.0 for both the predictions.

5.8 Another example result. The Recall@1 and StoryRecall@1 for $t_1$ and are 4.5% and 27.2% while for $t_{1,10,15}$ are 10% and 30% respectively. MAE for this example is 3.0 for $t_1$ as the model predicted 26th image as $\langle EOS \rangle$ instead of the 23rd image as in ground truth, while it is 2.0 for $t_{1,10,15}$.

6.1 Example story from our AESOP dataset with title and genres. The narrative is interesting, coherent and follows a clear causal arc with introduction and a moral at the end. The visual depiction of the story, including the changes in the expression of the characters, shows clear coherence and supports the narrative.

6.2 All the clip–art objects present in the dataset along with the four backgrounds are shown for reference. The objects are all scaled to a uniform size for display (This causes blurring of small objects like bee). The actual sizes depend on what the object is and is different than what is shown here.

6.3 Top half of the data collection interface showing the instructions to workers followed by some examples of good stories. Important and necessary instructions are highlighted in red. Otherwise, constraints are limited and the workers are asked to be creative.

6.4 Remaining part of the data collection interface showing a preview of the story in its current form (followed by canvas to create visual panel and then space to provide story text. To the right of the canvas are all the clip–arts split into categories that can be dragged to the canvas.

6.5 Once the stories are written, workers are asked to select the genres for the story they wrote from a list of predefined genres. They are also asked to provide additional comments if available.

6.6 Distribution of objects in visual panels across the entire dataset for Animals (Top), Large objects (Bottom).

6.7 Distribution of objects in visual panels across the entire dataset for Small objects (Top) and humans (Bottom).

6.8 Example stories from AESOP dataset with title and genres. The changes in location, pose and expression of objects align with the events in the story. Also events in the story have clear causality and coherence with a diverse set of backgrounds, poses, scenes and story arcs.
6.9 Example stories from AESOP dataset with title and genres. The short and interesting titles and the genres indicating the entire emotional arc in the story are useful for emotional perception and controlled storytelling tasks. They could also be provided as additional input to the model to condition the generation of missing panel or even entire stories.

6.10 More examples from AESOP dataset. we have stories that talk about clones, normal day of cooking, outing with mom and even a funny story with the title ‘Dumb and Dumber’.

6.11 More examples from AESOP dataset where objects are used creatively for purposes that does not usually define the object such as CD for wheels, seashell for masks and so on.

6.12 An example story from VIST (top) and AESOP (bottom) with two consecutive panels highlighted in Blue. Swapping the highlighted panels in VIST gives a story that is indistinguishable from the original showing lack of causality and coherence. In our dataset, swapping these panels would lead to a meaningless story.

6.13 AESOP model architecture containing a Text and Panel Encoders, followed by cross-modal attention and hierarchical decoders to generate a visual panel. (Zoom in for details)

6.14 Examples of Assistant Illustrator result by Ground truth, Human Baseline, Proposed model and Unimodal are shown.

6.15 Examples of Assistant Writer result by Ground truth, Human Baseline, Proposed model and Unimodal are shown.
6.16 Examples of Assistant Illustrator for last panel generation result by **Ground truth, Human Baseline, Proposed model** and **Unimodal** where the proposed model successfully generated relevant and coherent visual panels. **Analysis:** **Left:** The generated visual panel removes the doll, retains the woman and changes her expression correctly. Note that the human baseline and ground truth do not show clustered poses and hence looks more realistic while the generated visual panel has predicted the closest pose from the 20 possible poses (during training ground truth and input poses are clustered but general poses are shown here to retain the realism in data). Unimodal as expected does not know to remove the doll or change expressions indicating that the proposed model takes text into account. **Right:** The story starts at Alice’s house but ends at Ryan’s place. The ground truth story does not have a change of scene in the third panel. The human baseline however, correctly captures the change of scene. In the proposed model’s generation, the scene change from park to a house, Alice’s presence and her angry expression are all captured perfectly when compared with ground truth. However, the model misses that Ryan is lying on the couch.

6.17 Examples of Assistant Illustrator for last panel generation result by **Ground truth, Human Baseline, Proposed model** and **Unimodal** where the proposed model successfully generated relevant and coherent visual panels. **Analysis:** **Left:** The generated visual panel is close to ground truth as well as human baselines. It retains the sand, but relieves the chipmunk and has similar pose and orientation to human baseline for Ryan and the dog. Except for the turtle that seems to be on top of Ryan, the overall scene is relevant to the text, coherent to previous visual panels and depicts a meaningful scene. **Right:** The proposed model captures the change in expressions and relative locations correctly similar to ground truth or human baseline making for a reasonable illustration of the story. Moreover the model has learned to not vary the position of objects that are still such as the monkey–bars, bush, and tree.
6.18 Examples of Assistant Illustrator for last panel generation result by Ground truth, Human Baseline, Proposed model and Unimodal where the proposed model failed to generate relevant and coherent visual panels. **Analysis:** **Left:** The ground truth illustration for the last panel visualizes the Jared and dog before the dog goes to the other room while the human baseline visualizes them in another room. The proposed model also takes them to another room but lack of any details on what the room is in the story makes it difficult for the model to place them in reasonable locations. However, the model still got all the relevant objects such as couch and fireplace for the living room and dog, Jared, hamburger and the soda for the story. **Right:** The model illustrates falling down but did not capture where exactly Jane falls down which should have been near the pool. Predicting each of the attributes is a separate task in itself (e.g. spatial reasoning in abstract scenes formulated as a separate task in [73]), making the overall task complex.

6.19 Examples of Assistant Writer for last last panel generation result by Ground truth, Human Baseline, Proposed model and Unimodal where the proposed model successfully generated reasonable text to complete the story. **Analysis:** **Left:** The proposed model scored equally for relevance and meaningfulness against human baseline while the human baseline won against coherence. The proposed model’s generation won in all metrics against the unimodal GPT2 model. The generated text is a reasonable ending for the current story. **Right:** In this example the generated text by the proposed model captures the content with high relevance and coherence to the rest of the story achieving higher scores in the human evaluation.

6.20 Examples of Assistant Writer result for last panel generation by Ground truth, Human Baseline, Proposed model and Unimodal where the proposed model generated irrelevant or incoherent text to end the story. **Analysis:** **Left:** The proposed model’s generation shows how it loses on coherence while trying to be relevant to the corresponding visual panel. GPT–2 generates much more coherent text and keeping with the context of the other text panels makes it preferable. **Right:** This is another example of incoherent text generated by the proposed model. while the objects are perceived as we would like, the text is not comprehensible. Initializing the text parts of the model with pre–trained language models as pointed in future work would help overcome this limitation.

7.1 Example retrieved image by the proposed **Cross-Modal Coherence Model** (right) vs **Cross-Modal Coherence Agnostic Model** (left) for input caption (top).
7.2 Example image-text pairs from CITE++ (a) and Clue (b) datasets. Image-text pair on the left has relations Expansion, Elaboration and Temporal$\geq t$ while the one on the right has relations Action as Visible.

7.3 Confidence function $\eta_{i,c}$ with different $\lambda$.

7.4 Correct image rank vs. the difference between the similarities of the top 2 retrieved images on CITE++ validation set.

7.5 On CITE++ dataset, different values for $\lambda_{cls}$ vs MedR to determine the best value for $\lambda_{cls}$ as 0.1 (Left); Text sequence length vs MedR to determine the best value for maximum sequence length as 200 (Right).

7.6 Comparison MedR between baseline, CMCA and different CMCM variants; as well as the comparison between the same model with and without selective similarity refinement. Left: CITE++ dataset. Right: Clue dataset.

7.7 The ground truth image (Left) and the top 5 retrieved images by the CMCM and the CMCA models for two examples. The coherence relation (in blue) and caption are given above the images. The image-text pair in example (a) has Action relation while in example (b) has Visible relation. In example (a) the CMCM model leverages the Action coherence relation to retrieve images that depict some action in the top 5. Similarly in example (c) images retrieved by proposed our model with CAM retrieves images that depict the result of a process as given by the Temporal$\geq t$ relation, whereas the agnostic model shows images that depict action in progress.

7.8 Attention weights for CMCM and CMCA models for example (a) and (b) in Figure 7.7.

7.9 Example input text, ground truth image, ground truth image-text coherence relation and the top 4 retrieved images by the proposed CMCM algorithm and CMCA for comparison on the CITE++ dataset.

7.10 Example input text, ground truth image, ground truth image-text coherence relation and the top 4 retrieved images by the proposed CMCM algorithm and CMCA for comparison on the Clue dataset.
CHAPTER 1
INTRODUCTION

Communication is integral to society and survival. Human communication is inherently creative and multimodal. Even simple actions such as pointing and gesticling, hard for animals to comprehend, is easily understandable to humans around the world across regions, languages and borders. Even prelinguistic infants use and understand gestures [1] and they find it natural and transparent. Visual media from cave paintings and hieroglyphs to present day visual aids play a significant role in assisting in communication. What started off as mere pointing and pantemiming evolved into the multimodal notion of communication in the current world with visual and textual media coming together to communicate useful information.

For example movies, news and other televised media are mainly visual modalities, with associated audio from characters and possibly text (subtitles) as well to assist with communication. The visual scenes and the linguistic dialogues combine to tell a story.

Books, lectures and presentations, even if intended to convey something factual are presented with example narratives to make it easy for readers to understand. Books that do not tell a story or lectures that are mere descriptions of facts are uninteresting and disengaging.

There is an unprecedented wealth of multimedia data (image, video and text) on the web which stems from the availability of accessible imaging devices (e.g., cell phone and tablets) and the avid use of social media. Day to day communication through social media often involve sharing images, emojis [2] and other visual media [3] etc. associated with some text.
1.1 Why Study Stories?

Story and storytelling are simultaneously cognitive processes and products of cognition. Story is both art and quotidian, centripetal and centrifugal, running deep and wide through the human psyche. There is an abiding recognition that existence is inherently storied. Life is pregnant with stories. Quite possibly, it is the principal way of understanding the lived world. Story is central to human understanding—it makes life livable, because without a story, there is no identity, no self, no other [4].

Storytelling is integral to human experience. Starting from when we are very young, stories help shape our understanding of the world around us, and the people that inhabit it. Through stories, we encode a wide range of shared knowledge, including common sense physics, cause and effect, human psychology, and morality [5]. Storytelling and story comprehension are closely linked in that both involve the construction of rich mental models, comprising scenes, inanimate objects and their properties, as well as characters and their intentions [6]. Consequently, stories are crucial to mental development in humans. We postulate that machine intelligence requires comparable skills, particularly when interacting with people.

Storytelling itself is one of the oldest known human activities [7], providing a way to educate, preserve culture, instill morals, and share advice; focusing AI research towards this task therefore has the potential to bring about more humanlike intelligence and understanding.

1.2 Limitations of Prior Work

Availability of rich multimedia data along with recent algorithmic developments in neural architectures has resulted in the wealth of multi-modal image/video-text approaches. Typical problems such as language grounding of vision is generally treated as learning a *one-to-one* correspondence between text segments and associated images, such as image
captioning [8], image retrieval [9] or learning a joint embedding space [10, 11]. These techniques and datasets like [12] test the ability of a model to extract factual information about an image. However, text and images in the real-world usually co-occur to tell a unified story [13], such as in news articles, blogs, social network posts, comics, and movies. In order to be applicable in the real-world, a learned system must be able to comprehend stories that involves learning beyond simple factoid extraction [14].

1.3 Proposed Approach

We start by showing that existing datasets and generic vision-language tasks do not demand the understanding of factors such as coherence and causality that are important for story comprehension by qualitative and quantitative evaluation.

We propose Story Illustration, automatic illustration of an input story with a sequence of images, as a measure of story comprehension imitating how humans visualize when they read or comprehend a story. We develop an end-to-end trained neural module with explicit entity–grid based coherence modelling that is able to illustrate a story with clear understanding of coherence and co-reference resolution.

We then extend the Story Illustration to a more generalized Many-to-Many Story Illustration formulation and create a new dataset and a novel machine translation approach to story illustration. Our model is shown to generalize well and achieve high scores in creating coherent illustrations by virtue of its explicit causal decoding.

In our works, we observe that generation is a much closer imitation of the human visualization process than retrieval. Moreover, the existing datasets primarily capture the perceptive process associated with comprehension rather than the creative process associated with storytelling. We create AESOP, a new dataset that captures the creative process associated with visual storytelling. Using AESOP, we propose foundational storytelling tasks that are generative variants of story cloze tests, to better measure the creative and causal reasoning ability required for visual storytelling. We develop a generalized story
completion framework that models multimodal stories as the co-evolution of visual and textual concepts. Our dataset and model treats images as a composition of objects and related concepts making it capable of plugging them in different scenarios to generate new illustrations.

Additionally we address the limitation of universal characterization of all image–text pairs as captions by showing that coherence relations that characterize the different ways by which an image–text pair can be related improves the performance of generic text to image retrieval algorithms. We propose a Coherence Aware Module that enhances retrieval models with the ability to model coherence relations on a new dataset.

1.4 Summary of Contributions

Datasets:
We show limitations of existing datasets and their applicability for multimodal story comprehension. We identify factors associated with visual storytelling and propose novel datasets that overcome limitations of existing ones. We also assess the performance of the newly proposed datasets and compare them with existing ones in terms of quality, robustness and capabilities relevant to multimodal story comprehension.

Tasks:
We show that existing tasks do not demand true multimodal comprehension from AI systems. We show that simple yet effective neural models that when trained on proposed tasks perform better than a complex model trained on simpler and ill-defined tasks. We ensure that proposed tasks demand that models comprehend and model important factors associated with visual storytelling to perform well. Moreover, proposed tasks require models to be truly multimodal by creating and comprehending both the modalities simultaneously.

Neural Methods:
We model textual coherence in narratives to better comprehend stories. We propose simple machine translation based story comprehension systems that treat visual and textual
sequences as two different languages conveying common context. In a follow up work, we model generative and creative composition of stories by encoding the co-evolution of visual and textual concepts. We also show that modelling cross modal coherence relations improve performance of agnostic models for text to image retrieval.

1.5 List of Relevant Publications

* indicates the publication is under review and included in this thesis.

- **Ravi, H.**, Vithlani, P., Modi, A., Kapadia, M., Visualize Your Story: Laying the Foundation for Coherent Many to Many Story Illustration, **Under Review** *


CHAPTER 2
BACKGROUND AND PRIOR WORK

Recent advancements in deep learning research have led the fields of Computer Vision (CV) and Natural Language Processing (NLP) to see significant progress in several tasks. Independent of NLP, computer vision has achieved prominent improvements in tasks such as visual content classification [15], object detection [16], segmentation [17], etc., using self-supervision [18] or large annotated datasets. Similarly, independent from computer vision, NLP has seen a surge of interest in solving multiple tasks at once with unsupervised pretraining of language models ([19, 20]) using large unlabeled corpora. Other tasks include text generation [21] and narrative understanding [22]. There is also interest in solving challenges that combine linguistic and visual information from these traditionally independent fields. The methods which address the challenge of integration must provide complete understanding of visual or textual content as well as be able to semantically ground context in both the modalities. With the advent of large and comprehensive datasets like ImageNet [23], MSCOCO [12], Flickr [24, 25], CIFAR [26], etc. and efficient automatic deep feature learning methods such as Convolutional Neural Networks (CNN) [27] and Recurrent Neural Networks (RNN) [28] vision–language research has seen immense developments. The associated datasets, tasks and techniques usually fall into one of the following relevant categories.

• **Language Generation/Retrieval For Visual Input**: Most of vision language integration research formulates visual scene understanding by treating an image as input and information about the image as in natural language as output. Common tasks include image captioning [29], visual question answering [30], Visual entailment and image recognition [31, 32], dense captioning [33] and others. Many techniques such as video captioning [34] use a single video clip as visual input instead of an image.
• **Visual Scene Generation/Retrieval For Language Input:** Some techniques treat natural language text description as input and an image as output to visually ground textual context. Examples include text to image generation [35] and retrieval [36]. Text to video retrieval [37, 38, 39] is a common research problem while video generation [40] is gaining interest.

• **Joint Understanding Of Vision and Language:** Joint representation learning techniques combine some of both image/video–to–text and text–to–image/video tasks by embedding both the modalities in a joint embedding space. These techniques can simultaneously solve image/video–to–text retrieval, text–to–image/video retrieval and image/video–text alignment [10, 11, 41] tasks.

• **Comprehension Of Multimodal Stories:** Contrary to other categories, visual storytelling [42, 43] deals with aligning a sequence of images with a natural language story to extend the semantics beyond a single image or text for narrative understanding. They differ from generic tasks by extending the reasoning from a single image–text pair to a sequence of images and text. Note that even though videos are a sequence of frames, they are usually aligned as a whole with corresponding text in video–text understanding techniques.

We discuss some relevant works in each category in detail below, identify their limitations and why they are insufficient to develop systems that can create and comprehend multimodal stories.

2.1 **Language Generation/Retrieval For Visual Input**

Vision language integration gained a lot of traction with the introduction of large curated datasets like MSCOCO [12] and MPII [44, 45] for image/video captioning.
2.1.1 Image to Text

Caption generation for images requires a model to understand the contents of an image and describe the same in detail using natural language. This is a well-studied problem in the vision community [46, 47, 48, 49, 50, 51, 52, 53]. Most recent works use some form of RNNs to generate the captions word by word given the encoding of the image generated using CNNs. Particularly, He et. al. [54] use Part of Speech (POS) tags from image descriptions, while Jia et. al. [49] utilized semantic information from images to guide Long Short Term Memory (LSTM) to generate meaningful descriptions. Vinyals et. al. [55] and Chen et. al. [56] used CNN-based image encoders and an RNN and bidirectional RNN for decoding, respectively. Attention-based neural networks are also popular and allow the model to focus on specific regions when generating individual words [52]. Johnson et. al. [57] proposed a convolutional localization network to create dense captions for an image in a single forward pass. Captions, in general, have been defined as a descriptive piece of text that contain information about the objects and scene in the image. [29, 58, 59, 33, 60].

2.1.2 Video to Text

Videos provide the necessary temporal information to model visual scene understanding. Datasets for video-text alignment span movies [45, 44] to instructional videos [61, 62]. Venugopalan et. al. [63] use the MPII movie description dataset [45] and MSVD dataset [44], to train a video to text, sequence to sequence network. Kaufman et. al. [38] use temporal tessellation for video analysis and captioning. Shen et. al. [64] introduce a weakly supervised strategy that first understands region level semantics of frames in an video and then generate multi level sentences for region sequences followed by dense captions. Wang et. al. [65] propose a encoder–decoder–reconstructor architecture as additional objective to ground the generated captions. In [66], the authors propose a hierarchical reinforcement learning framework for captioning to address the limitations of simple sequence–to–
sequence networks.

2.1.3 Beyond Descriptive Captions

A major limitation of generic image to text retrieval or generation techniques in the literature is that, the text that is associated with the image often provides only factual information about the image without adding any new knowledge [3]. There has been increasing interest in modelling the subjective attributes of image–text representation learning by associating an emotion label [67, 68], hashtags [69], personalization [70] and cross–modal coherence labels [13] with image-text pairs. These works extend captioning beyond descriptive text to controllable generation, subjective personality based captions etc. Other works along the same line include [71, 72]. Images are composed by clip–art objects in [73, 74, 75, 76] where the aim is to model image–text relationship from the perspective of abstract visual reasoning.

2.2 Visual Scene Generation/Retrieval For Language Input

Language to vision systems use natural language text as input and some form of visual information such as image/video as output.

2.2.1 Text to Image

Image retrieval is considered to be the reverse of caption generation and used in several multimodal NLP tasks and applications. [77] extract syntactic relations from captions for indexing and retrieving photographs of crime scenes. [78] use image retrieval as a testbed for learning spatial relationships between image regions using Visual Dependency Representations. Several works have shown that including images in information retrieval tasks such as document retrieval can improve the performance of the models [79, 80]. Most recent visual dialogue systems include image retrieval models to present images with text in response to user’s needs and to better fulfill the dialogue goals [81].
Reed et. al. [82] proposed a Generative Adversarial Network to synthesize images from descriptive text. Text to Image generation, in contrast to retrieval generated the entire image pixel–wise from scratch. This task is much more complex and has gained a lot of attraction currently. Dong et. al. [83] use Generative Adversarial Networks (GAN) to generate an image that matches a description while retaining the visual semantics of another similar image. In [84], the authors use a bidirectional RNN with attention mechanism over text to iteratively draw on a canvas using variational encoding. Xu et. al. [85] use Attention as part of GANs for text to image synthesis while Qiao et. al. [86] propose a text-to-image-to-text framework for image synthesis to be regularized using caption generation. Similarly, [87] propose cross modal contrastive GAN for text to image synthesis to achieve state of the art results.

Zitnick et. al. [75] propose the abstract scenes dataset for reasoning over abstract visual concepts and formulate text–to–image generation as drag and drop of objects and attributes rather than generating pixel values like described above. Kim et. al. [76] build on top of that to propose a collaborative framework where two systems interact with each other to iteratively generate visual scenes. In [74], the authors use the abstract scenes dataset [75] to generate the scenes directly from captions using sequential convolutional networks.

2.2.2 Text to Video

Text to video retrieval is also a commonly studied problem and usually concentrates on aligning the video content with textual context. Traditional retrieval methods are content based [88] where hand crafted features are extracted from processed segments of the video, clustered and then indexed for retrieval. Then, Dong et. al. [89] propose deep representation learning methods for text–to–video retrieval using sequential networks for encoding. Wu et. al. [90] proposes a deep hashing framework for large scale retrieval of videos. In [91], the authors propose an efficient global–local matching technique for text to video retrieval.
Text to video generation has also gained traction with the popularity of GANs and Variational Auto Encoders (VAE). Condrick et. al. [92] first used GANs for video generation. Sync-DRAW [93] combined VAEs with RNNs to generate videos trained on Bouncing MNIST [94] and KTH [95] datasets. Li et. al. [96] combine VAEs and GANs while Den et. al. [97] propose convolutional GANs for video synthesis. Balaji et. al. propose conditional GANs with discriminative filters for text to video generation.

2.3 Joint Understanding Of Vision and Language

Joint understanding of vision and language unlike specific vision to language or language to vision systems encode both the modalities on a common embedding space to perform cross–modal retrieval or generation in both directions. For example, Wang et. al. [98] proposed structure preserving and bidirectional ranking constraints, while Zhou et. al. [99] formulated a Gaussian visual-semantic embedding; Gong et. al. [100] used a multi-view version of Canonical Correlation Analysis (CCA) for joint representation of phrases and images. Vendrov et. al. [101] proposed to leverage intuition that a correct caption-image pair can be considered as ordered, with the caption being a more abstract representation of the image. This effectively encoded entailment relationship between images and captions. Evaluations for this method produced state-of-the-art results for image retrieval and caption generation.

Some recent works [11, 102, 41, 103] have proposed large multimodal pretraining networks based on the Transformers [104] architecture that obtain state of the art results on more than one specific image–text understanding task.

Large pretrained multimodal transformers [105, 62, 106, 107, 108] have been proposed for joint representation learning of videos and text. Most of these follow suit of multimodal transformers like [41] and propose multiple transformer encoders that embed the visual and textual modalities either jointly or separately and then learn a joint representation objective.
2.4 Limitations of Vision-Language Techniques

Through extensive research on various complex deep learning frameworks and objectives, vision to language tasks such as captioning, recognition and others have seen remarkable progress. Currently, image captioning systems trained on standardized datasets like MSCOCO perform close to humans as evaluated using text generation metrics and human evaluation. However, these datasets and associated techniques have major limitations and are insufficient to model narrative understanding.

The techniques described above treat captions as descriptive piece of factual information about an image. In contrast, regular image–text occurrence in social media (e.g. Twitter, Facebook, Instagram), books and literature, or any form of communication uses text that is narrative and abstract [42, 3]. Though controlled caption generation [67] hopes to capture user intent (subjective), the captions are still grounded to a specific emotion and largely descriptive. This is in stark contrast to day-to-day human discourse.

Another limitation is that all these works portray visual scene understanding as a static problem where text is used to describe a single static image whereas Cognitive and Neuroscience literature [109, 5, 110] suggests that visual perception is an abstract and temporal process. Though Video and Text systems work with visual information that has temporal context, the associated caption usually describes a particular moment in the video clip rather than evolving with the video. Also, all the techniques encode the video clip as a whole and often separately from text and then align the encoded representations.

Also, all these techniques perform text generation from visual input or the reverse whereas the possibility of co-creation of both text and images/videos is not studied. In reality textual and visual information often evolve together to tell a story [13, 42]. Beyond understanding simple objects and concrete scenes lies interpreting causal structure; making sense of visual input to tie disparate moments together as they give rise to a cohesive narrative of events through time. This requires moving from reasoning about single images –
static moments, devoid of context – to sequences of images that depict events as they occur and change [42].

2.5 Narrative Understanding Of Text Only Stories

Narrative understanding has been extensively studied in natural language [111, 112]. Some works focus on story understanding from the perspective of learning scripts [113] while [114, 115, 116, 117] perform unsupervised learning of event schemas and narrative chains from stories. Recent datasets such as ROCstories [118] have accelerated deep learning for story comprehension research via datasets, standardized cloze style tasks and metrics. Towards this, [119] proposed a hierarchical plot plus story generator while [120] use common sense knowledge base conceptNet for story comprehension. Similarly [121] proposed a scene graph approach to story generation while [122] extended the dataset to provide causal event annotation to study causality in stories. Other similar works include [123, 124, 125, 126]. Although these works have extensively studied natural language stories, visualization of the stories as a function of a model’s ability to comprehend stories has not been investigated provided that human communication and perception are inherently multimodal.

2.6 Comprehension Of Multimodal Stories

Visual Storytelling was introduced in [42] with the VIsual StoryTelling (VIST) dataset. It contains sequences of five images, aligned with one sentence describing each image forming a story. A CNN-RNN baseline was shown to create meaningful stories. Following that, [127] proposed to learn a joint embedding space to address the huge variance in story image sequences of the VIST dataset. [128] tackled album summarization and visual storytelling together while [129, 130] leverage adversarial techniques to learn storytelling rewards. [131] proposed a manager-worker network to explicitly handle exposure bias and train on a Reinforcement Learning objective. Similarly, [132] propose composite rewards
both trained on reinforcement learning objectives. Yu et. al. leverage large pretrained transformer models and propose a transitional adaptation step to better learn common context between vision and language gaining state of the art performance in VIST [42] and LSMDC [133].

2.7 Limitations of Story Comprehension Systems

Though there have been a lot of works focusing on natural language narrative understanding, the lack of visual modality largely impacts its applicability to the real world. As outlined in chapter 1 human communication is predominantly multimodal. Visual narratives are largely unexplored and systems that work for natural language narratives do not directly extend to situations with visual information.

Through Visual storytelling [42], vision language intersection scratched the surface of multimodal narrative understanding. However, the dataset and associated tasks and techniques have major limitations. Moreover, comprehending complex multimodal narratives requires modelling the co–evolution of both the modalities rather than treating it as a cross–modal retrieval or generation task. The techniques proposed to address visual storytelling in VIST, model relationship between sentences in a story and then map each sentence to one image as constrained by the dataset, restricting its applicability to model a general coherent narrative [3]. Consequently, a trained model is required only to produce a “feasible” text for a given image sequence without modelling intent or the overall narrative structure in the story.

Through this thesis, we propose novel datasets, tasks and neural methods to address many of the limitations highlighted in this chapter to take further steps towards creating AI systems that can comprehend multimodal narratives.
CHAPTER 3
MULTIMODAL STORY COMPREHENSION - DATASETS, TASKS AND NEURAL METHODS

This dissertation is based on the following main objectives and contributions:

Datasets:
Show limitations of existing datasets and their applicability for multimodal story comprehension (chapter 4, chapter 5 and chapter 7). Identify factors associated with visual storytelling and requirements for a good dataset (chapter 6). Propose novel datasets that overcome limitations of existing ones and also satisfy the requirements for demanding true multimodal comprehension from systems trained on them (chapter 5 and chapter 6). Assess the performance of the newly proposed datasets and compare them with existing ones in terms of quality, robustness and capabilities relevant to multimodal story comprehension (chapter 6).

Tasks:
Show that existing tasks do not demand true multimodal comprehension from AI systems. Also, tasks are cross-modal and ignore the co-evolution of visual and textual modalities (chapter 4 and chapter 5). Model coherence explicitly and propose simple yet effective neural models that when trained on proposed tasks perform better than a complex model trained on simpler and ill-defined tasks (chapter 5 and chapter 7). Ensure that proposed tasks demand that models comprehend and model important factors associated with visual storytelling to perform well. Moreover, tasks must be truly multimodal by creating and comprehending both the modalities simultaneously (chapter 6).

Neural Methods:
Model textual coherence in narratives to better comprehend stories (chapter 4). Propose simple machine translation based story comprehension systems that treat visual and tex-
tual sequences as two different languages conveying common context (chapter 5). Model generative and creative composition of stories by encoding the coevolution of visual and textual concepts chapter 6. Propose and model cross modal coherence relations to improve performance of agnostic models for text to image retrieval chapter 7.

The dissertation starts by providing introduction, motivation and applications of multimodal story comprehension in chapter 1. chapter 2 discusses related works while chapter 4 to chapter 7 are based on papers from refereed publications. Finally chapter 8 provides conclusion and discusses future possibilities of this dissertation.

### 3.1 Illustration as a Measure of Comprehension

![Figure 3.1: Illustration showing the story illustration task. Input is surrounded by a red box while output is in green box. The blue arrows indicate the reasoning pathways for the proposed model and tasks.](image)

Current vision–language integration approaches have started to address visual storytelling by proposing datasets (e.g., VIST [42]) and hierarchical language decoders that are able to generate multiple sentences [134], or paragraph descriptions, forming stories [135]. While some limited success has been shown, most of these approaches attempt to go in a forward direction, producing a multi-sentence description for an image [134], video [136], or image sequence [42]. Moreover, visual storytelling as a task does not demand actual comprehension of the input story from a model, to perform well.

In chapter 4 we propose *Story Illustration*, the task of automatically retrieving a se-
quence of images to illustrate a given input natural language story. This task can be thought of as imitating how humans automatically visualize the events and characters in a story while comprehending the same. An example input, output and reasoning pathways is shown in Figure 3.1. We describe an end-to-end network for visual illustration by building a neural architecture that encodes the input using hierarchical two-level sentence-story GRU, combined with an encoding of coherence, and then sequentially decodes the predicted feature representation into a sequence of images. We show state of the art performance with comprehensive evaluation.

3.2 Modelling Coherence and Many–to–Many Mapping

Figure 3.2: Illustration showing the Many–to–Many story illustration task. Input is surrounded by a red box while output is in green box. The blue arrows indicate the reasoning pathways for the proposed model and tasks.

Though our CNSI model for story illustration proposed in chapter 4 performs well, we observe that neither the dataset [42] nor the model can be generalized to situations where images and text do not have one–to–one correspondence. On the contrary, in storytelling [42], text decoders can generate stories of any length in an open ended manner. This constraint limits the applicability of these techniques to real-world applications. The reasoning pathways shown in Figure 3.1 also ignore the coherence between images in the illustration.

chapter 5 proposes a more generalized task called Many to Many Story Illustration that does not assume one–to–one correspondence between sentences in the input story and
images in the illustration. To model this task, we propose a novel many–to–many dataset using the MPII dataset [45]. We then treat the input story and output illustration as two different languages conveying the same context and use machine translation type encoder–decoder. This simple model is generalizable and performs better than previous approaches.

3.3 Multimodal Creative Composition

![Illustration showing the AESOP Assistant tasks. Input is surrounded by a red box while output is in green box. The blue arrows indicate the reasoning pathways for the proposed model and tasks.](image)

Figure 3.3: Illustration showing the AESOP Assistant tasks. Input is surrounded by a red box while output is in green box. The blue arrows indicate the reasoning pathways for the proposed model and tasks.

One of the major limitations of current approaches in addressing visual storytelling is that datasets do not capture the creative process associated with storytelling. Storytelling is a creative cognitive process [6]. Though the task of Story Illustration (chapter 4 and chapter 5) is a reasonable measure of story comprehension, associated datasets like VIST [42, 137] lack coherence and diversity. Story Illustration as a task is formulated as the retrieval of a sequence of images from a large dataset. Human visualization process is better modelled as generative with the ability to compose visualizations using abstract concepts.

In chapter 6, we propose AESOP: a novel dataset that captures the creative process associated with visual storytelling. Inspired by [76, 75, 74], our dataset employs abstract visual scenes, with a broad set of choices for objects and attributes needed for visual storytelling essentially enabling the modelling of images as a set of concepts. We propose novel story comprehension tasks on AESOP that demands multimodal, abstract, creative
and causal reasoning ability from visual storytelling systems. The reasoning pathways shown in Figure 3.3 models the co-evolution of both the modalities in contrast to previous setups. Further, we propose a novel generalized story comprehension framework that models stories in our dataset as the co-evolution of visual and textual concepts.

### 3.4 Image–Text Coherence

Most of our work on story comprehension deals with a sequence of images and text while even a single image–text pair can convey a unified story [13]. In chapter 7, we study the role of coherence relations in text-to-image retrieval. We hypothesize that bringing in coherence relations [13] into the retrieval process, in contrast to personalities defined in [138], should better improve the performance of text-to-image retrieval in a more generalizable way. Our proposed framework introduces a *Coherence Aware Module* that learns to predict coherence relations during training, and the same module is also applied during testing to further improve the retrieval performance. This module helps the retrieval model focus on the intent of the users and the potential effects of image–text combination.
CHAPTER 4

ILLUSTRATION AS A MEASURE OF COMPREHENSION

Visualization is an important and involuntary process by which humans comprehend stories. Imagine reading a novel that does not have any explicit illustrations. We still tend to visualize what is being read in real time to disambiguate details and make inferences [139]. Visualization is a process by which we compose our own version of what we understand while reading a story. Composing is critical to thought processes because it is a process which actively engages the learner in constructing meaning, and developing, relating and expressing ideas. Comprehending is critical because it requires the learner to reconstruct the structure and meaning of ideas expressed by another writer. To possess an idea that one is reading about requires competence in regenerating the idea, competence in learning how to compose the ideas of another. Thus both comprehending and composing seem basic reflections of the same cognitive process [140].

Humans draw upon deep world knowledge, grounded in visual-linguistic stories and experience that we’ve accumulated from a young age. Therefore, it is likely worthwhile to similarly ground machine comprehension and synthesis of stories in the visual world. Much of the current work on joint understanding of vision and language hinges on learning to describe factual information about objects and scenes in an image. Typical problems include image captioning [57], natural language-based image retrieval [82], and joint embedding of text and images to understand the relationship and be able to translate between the two modalities [98]. However, most of these formulations (??) assume atomic image-sentence pairings both at training and test time. This makes it difficult to apply them for storytelling tasks where sentences are implicitly inter-related in a narrative. Recent approaches have started to address these challenges by proposing datasets (e.g., VIST [42]) and hierarchical language decoders that are able to generate multiple sentences [134], or
The fireworks this year were amazing. Some were like colorful star-flowers falling from fire bursts. Some went multicolored from green to red. Many exploded from the ground making a peacock tail of fire. Brilliant colors of light made for a beautiful night.

Figure 4.1: Two image sequences visualize the given story (top). The images predicted by the proposed sequential model with coherence (bottom) demonstrate higher consistency and better alignment with the story than the images retrieved independently sentence-by-sentence (middle).

A key difference with story captioning is that images tend to be more expressive (e.g., an image is worth a thousand words), making it challenging to produce a coherent sequence of illustrations. We also propose a neural method for retrieving a sequence of illustrative images which correspond to a narrative passage of text. Figure 4.1 illustrates the problem and our solution (CNSI) for a given input passage (top). The images in the middle row are predicted by the baseline where a sentence-image similarity is learned [101] and images are independently retrieved for each sentence in the passage. The key benefit of encoding context, through hierarchical Gated Recurrent Units
GRU) [142], and coherence is illustrated by images in the bottom row. In fact, only the first sentence in the story mentions the word ‘fireworks’, but all of the images in the output share that visual, highlighting that the references were correctly resolved. This example clearly illustrates the key benefits of our proposed model over most current research which mainly focuses on understanding the relationship between descriptions and corresponding images as atomic one-to-one relations. Even techniques ([141]) that work for sequential text generally consider large pieces of text that have more descriptive sentences. Information flow in such text is mostly restricted and parts of text tend to behave distinctly which does not necessarily require sequential modeling. Storytelling sequence of sentences however have coherence between sentences with distributed information that have to be modeled sequentially.

4.1 Story Illustration

Visual Storytelling is formulated as composing a natural language story that describes a sequence of input images [42]. Consequently, this models the perceptive process associated with storytelling i.e. a user tries to come up with a story based on what they perceive of the input sequence of images without an intended story to ground the composition. Moreover, the absence of an intended story, makes it harder to measure the comprehension ability of a model. Given how the dataset has been developed, we believe it is preferable to formulate the inverse problem of illustrating a natural language story with a sequence of images, as a measure of story comprehension.

Problem Formulation

We propose Story Illustration, the task of automatically retrieving a sequence of images to illustrate a given input natural language story. This task can be thought of as imitating how humans automatically visualize the events and characters in a story while comprehending the same. Since the stories in VIST are obtained for an input sequence of images, the
inverse problem of retrieving the sequence of images given the story ensures there is an intended visualization that can be used as the gold standard. We develop the SIS to IIS retrieval algorithm as follows. Let $S = \{s_1, s_2, ..., s_n\}$ be a set of $n$ sentences (though in principal these can also be paragraphs or other natural text elements) that tell a narrative story and let $I = \{i_1, i_2, ..., i_n\}$ be the set of $n$ corresponding images from the dataset that best illustrate the input SIS. We consider $n$ to be equal to five in our paper to adhere with the VIST [42] dataset, however, the algorithm is general and is able to deal with arbitrary length sequences. There is, however, an implicit assumption that illustration is one-to-one, meaning that for every sentence (element of the story) we retrieve one illustration.

We develop an end-to-end network for visual illustration of a sequence of sentences forming a story. We refer to an input storytelling sequence of sentences as SIS (story-in-sequence) and the corresponding output visual summary of images as IIS (images-in-sequence). We achieve this by building a neural architecture that takes the form of an encoder-decoder, where sentences are encoded using hierarchical two-level sentence-story GRU, combined with an encoding of coherence, and then sequentially decoded using the predicted feature representation into consistent IIS. We optimize all parameters of our network in an end-to-end fashion with respect to the order embedding loss. The resultant model tries to sequentially translate each input sentence vector to a representative output image vector. The image closest to this output vector is then retrieved from a large dataset to illustrate the corresponding sentence.

Due to the ambiguity of the task, existing quantitative metrics, such as mAP [143], produce misleading results, as these metrics are computed based on the exact image IDs. For example, as shown in Figure 4.2, the precision goes to zero because predictions (in the bottom) consists of different images from the ground truth (middle). However, the predicted images are actually preferred, to ground truth, by Amazon Mechanical Turk (AMT) subjects in terms of better story visualization. To address this, we first perform a user study on AMT to evaluate the performance of the proposed architecture, in comparison to ground truth,
I choose to take black and white photos today. An old leaflet I found on the floor of the cemetery. I tried to read this but it was in a foreign language. These must have been wealthy considering the size of the their tombstones. And that concluded the day at the Cemetery.

The friends met up for drinks. It was a fun night out for everyone. The next morning everyone played instruments. Ron had played the guitar since he was a kid. That night everyone got ready to go out and do the same thing again.

Figure 4.2: Two samples of ground truth (middle) and our prediction (bottom) for the story (top). The bottom sequence of images win more votes from AMT workers for better visual illustration.

and existing baselines [101] of visual retrieval for isolated sentences, and ablated model without coherence. Results indicate that the proposed model outperforms the baseline, and users prefer image sequences with coherence. Additionally, we propose a new quantitative metric for this task based on the visual saliency of the retrieved images with respect to the ground truth images. We show this metric serves as a good proxy for measuring whether a predicted image can be considered a good visual illustration.

Contributions

Our core contribution is an end-to-end architecture for retrieving a sequence of illustrative images from a set of sentences, one for each sentence, forming a story. We model context between sentences using hierarchical two-level sentence-story GRU. Further, since it is natural to use references (e.g., direct: he/she/it and indirect: they both went there) within the story once actors and objects are defined, we also introduce a coherence vector to help with such reference resolution. Evaluation of the proposed architecture with a user study shows that the proposed algorithm performs better than the baseline in a comprehensive ablation study. Further, we introduce a new metric for this task that can better deal with ambiguities in the image selection.
4.2 Visual Storytelling Dataset

We use VIST dataset [42] for all of our experiments. To our knowledge this is the only dataset that consists of sequences of images with sequences of text descriptions that form narrative stories. The dataset consists of approximately 40,154 stories for training, with each story made of 5 sentences and corresponding set of 5 images. The sentences are unique for each story, but the sets of images are not unique. In fact, out of the 200,770 images (40,154 x 5) only 65,145 are unique. In addition, there is a test set (5,054 stories) and validation set. Note that SIS is present for all of the 40,154 stories while DII are present only for a subset of 28,000 stories. This is the case with the test set as well. Even though the dataset contains story image sequence repeats, this shortcoming actually represents a more realistic scenario and so we use the dataset as is.

4.3 Coherent Neural Story Illustration

We build an end-to-end network, as shown Figure 4.3, which encodes the input sentences and coherence and predicts encoding of the feature representation for corresponding illustrative images. We now explain each component of the network in detail.

4.3.1 Sentence Encoding

The process of sentence encoding is illustrated in Figure 4.3(b). Let each input sentence $s_j$ have $n_j$ words $\{w_1, w_2, ..., w_{n_j}\}$. The GRU RNN network in the first stage sequentially encodes $\{w_1, w_2, ..., w_{n_j}\}$, to generate feature vector $f_1(s_j)$ representing the $j$-th sentence. The corresponding image feature vector $g_1(i_j)$ is obtained by using a pre-trained VGG-16 model [144]. This entire network is trained on the MS-COCO dataset [145] so that $f_1(s_j)$ will be aligned with $g_1(i_j)$. Conceptually, this is to form initial sentence representations that are closer to image vector representations for each sentence in the story. This puts the text and image data in a common latent space which simplifies training of the full model. The
The family is celebrating the wedding with a reception.
Lots of relatives came from far away.
The new couple was very happy.
The reception was held next to the marina.
The cake was huge and beautiful.

Figure 4.3: The proposed approach for story visualization: (a) Shows modeling of isolated sentences for sentence encoding and parse tree extraction for coherence vector computation. (b) Uses encoded sentence vectors to train sentence-RNN using Order Embedding (OE) loss function [101]. (c) Uses encoded sentence and image vectors along with coherence vector to sequentially encode the input story. Story-RNN is trained using modified OE-loss function with On-line Negative Mining (ONM) for generating negative samples within a batch.
Order embedding loss (OE-Loss) function as defined in [101] is used to train this network. These encoded representations are then used to initialize the next part of the network that performs sequential story encoding in Figure 4.3(c).

### 4.3.2 Story encoding

The output of the sentence encoder is a sequence of individual sentence encodings: \( \{ f_1(s_1), f_1(s_2), \ldots, f_1(s_n) \} \). To encode the story structure among them, we pass this to a higher level network that encodes the sequential (from left to right) nature of the input story to produce vectors: \( \{ f_2(s_1), f_2(s_2), \ldots, f_2(s_n) \} \). The target vectors remain the same from above \( g_1(i_j) \).

A sequential, order embedding loss function is used to train this network to constrain \( f_2(s_j) \) to be as close as possible to \( g_1(i_j) \). The sentence encoding part is trained offline and is set as non-trainable while the the story encoder is training. This part of the network tries to minimize the cumulative loss over all sentence-image pairs in a story. This process is depicted as the sequential model in Figure 4.3 (c). For sequential story encoding, a three layered RNN with GRU cells [142] is applied to model the shared context between sentences in a story.

### 4.3.3 Coherence Model

Even though the GRU sequential model encodes the relationship between sentences, we also explicitly model co-references between sentences to further improve the ability of our model to capture story structure. We do this by making use of the coherence model proposed in [146]. The authors represent the coherence between sentences, within a piece of text, by a 64 dimensional coherence vector obtained from the parse tree associated with each sentence in the story. They use this vector as input to the final Fully Connected (FC) layer in their network, after zero padding to match dimensions of the input vectors. We, however, directly concatenate the vector with each sentence before the final FC layer. This is visualized in Figure 4.3(c).
4.3.4 Loss function

We use an order embedding loss function based on the one proposed in [101] for our network. The assumption is that a short textual description of an image is more abstract than the image itself. The description-image pair can therefore be considered as an ordered pair. Since SIS text is in general more abstract, we use a similar order embedding constraint based model. For ours, the cost of a sentence-image ordered pair violating the order is defined as:

\[ E(x, y) = \max ||0, (y - x)||^2 \]  

(4.1)

where \(E(x, y) = 0 \iff x \preceq y\) according to the reversed product order. If the order is not satisfied, then \(E(x, y)\) is positive. If we treat the sentence-image pair \((s_j, i_j)\) as a two level partial ordering, then we can define \(S(s_j, i_j)\) as follows:

\[ S(s_j, i_j) = -E(g_1(i_j), f_2(s_j)) \]

\[ = -\max ||0, (f_2(i_j) - g_1(s_j))||^2 \]  

(4.2)

where \(S(s_j, i_j)\) is the negative order violation penalty for a ground-truth sentence-image pair. The objective is then to maximize this for a ground truth pair relative to other pairs by a margin. Here, \(f_2()\) and \(g_1()\) are the SIS and IIS encoders as described in Sec subsection 4.3.2. The loss function to be minimized is then:

\[ c = \sum_{k=1}^{l} \sum_{j=1}^{5} \left( \sum_{s_{k,j}} \max \{0, \alpha - S(s_{k,j}, i_{k,j}) + S(s_{k,j}', i_{k,j})\} + \sum_{s_{k,j}'} \max \{0, \alpha - S(s_{k,j}', i_{k,j}) + S(s_{k,j}, i_{k,j}')\} \right) \]

(4.3)

where \(c\) is the cost for a batch with in size of \(l\). Index \(k\) iterates over each story within a batch while index \(j\) iterates over each positive ground truth sentence-image pair within each story. Given a batch of story sentence-image pairs, we apply Online Negative Mining
(ONM) to generate negative samples [101]. The negative samples for each ground-truth pair are taken from all other stories except the one in consideration. In other words, for a sample \((s_{1,1})\), the corresponding negative samples are \((s'_{k,j}, k \neq 1, j = \{1, 2, 3, 4, 5\})\). Also, \(j\) is chosen uniformly at random between the five indices for each negative story. Before each epoch, all the samples are arranged and shuffled carefully to avoid identical images occurring in different stories.

4.4 Experimental Setup

In this section, we discuss the entire experimental setup, training details, baselines and metrics used for evaluation of CNSI. Evaluating the performance of the algorithm for SIS to IIS is non-trivial as there may be multiple correct sequences of images that each can visually describe a given story. For example, Figure 4.4 shows a story that is visually represented by two sequences of images. It is hard to tell which one of the two visual depictions is the correct one as both sequences describe the story adequately. The VIST [42] dataset has many stories where visual coherence is not explicit (in terms of common objects or scenes throughout the story). Also, since stories are short, the possibility of a visually coherent object or scene being present is generally low. Hence, we resort to evaluation with a user study for a reference of what is correct. We then try to replicate the results using a visual saliency based quantitative metric. We first describe the training details and our baselines and then present the results.

4.4.1 Training Details

The proposed hierarchical GRU network with order embedding loss function can be completely seen in Figure 4.3(c). The sentence encoder is trained on MS-COCO dataset [145] using a joint image-sentence embedding formulation. We believe this ensures a good initial aligned representation for both modalities. The resultant vector for each sentence is given as input to the sequential model to encode the relationship between the sentence vec-
tors. The loss function is explained in Sec subsection 4.3.4; it calculates loss between the five encoded vectors and the corresponding five image vectors obtained from a pre-trained VGG–16 CNN network. For learning we use Adam optimizer \[147\] with a learning rate of 0.001. The batch size is 32 stories, a relatively low number to prevent repetition of stories or images in each batch.

Note that out of the 40, 149 stories present in the training dataset for VIST [42], there are only 16, 041 unique story image sequences. This means that multiple SIS can correspond to a single sequence of images. There is a high possibility that in a naïve implementation the order embedding loss function would get the same sequence of images as both positive and negative during the training. This is obviously undesirable. In addition, multiple stories can share different permutations of the same sequence of images causing same image to be seen as both positive and negative illustration for a sentence. Also, the algorithm performs prediction over the entire dataset for each time instant (sentence) during retrieval; therefore during training, we apply Online Negativing Mining to obtain the negative samples from dissimilar stories \(k \neq 1\) in Equation 4.3) from a disjoint set of instants \(j\) chosen uniformly at random in Equation 4.3).

The learning is set to run for 150 epochs. From observations, we find that after 130 epochs the loss value starts to saturate at around 3.0. During testing, each sequence of input sentences goes through the network and produces a sequence of image vectors. All images in the dataset go through the CNN part of the pre-trained baseline. Then, for each output vector from network, the image with the closest CNN feture vector is chosen as the predicted image.

We implement our networks and loss function in python using Tensorflow and Keras. The network is trained on the 40, 154 training set stories in the VIST [42] dataset. We remove stories that have broken URLs or images. We perform qualitative and quantitative evaluation on a subset of test set stories that have captions for all of the images. This reduces the number of test stories from 5,054 to 3,384. Retrieval is performed over this
Graduation day has finally arrived. All the students started filing in. He was thrilled to finally be graduation. Everyone posed for pictures outside the venue. And to cap off the day; we all hung out at the pool.

Figure 4.4: Two image sequences visualize a "graduation" story. AMT workers prefer the GT over BL, though both look similar.

entire set with 5,055 candidate images, *i.e.*, 3,384 stories with 5 images each in the test set have a total of 5,055 *unique* images. Each epoch takes approximately 250 seconds in a desktop with an NVIDIA Quadro K2200 GPU and CPU With 32 GB RAM and 1 TB Hard disk space. Prediction takes about 0.5 sec. per image.

4.4.2 Comparative Evaluation

We compare our approach with two baseline networks to analyze the different aspects of the proposed network.

**BL: Baseline Network.** For the baseline network, we use the one-to-one description to image retrieval algorithm proposed in [101]. The network is pre-trained on MS-COCO [145] and then trained on VIST [42] dataset. Each sentence in the SIS and image in the corresponding IIS are separated from their story to get 200,770 separate sentence-image pairs. We train on this and rearrange the test set similarly for evaluation. This experiment is to study the importance of sequential modeling of the input story.
**NSI: Network without Coherence.** In order to evaluate the effect of coherence on the performance of the network, we also consider the proposed network without the coherence vector. Since each sequence of images in the dataset has a different amount of coherence, the role of coherence in our algorithm needs to be assessed. The training procedure is identical to our full model.

### 4.4.3 Visual Saliency Based Metric

We observe that defining one visually correct sequence of images is not trivial in storytelling. As shown in Figure 4.4 (GT and BL), multiple visual summaries can describe a story without ambiguity. Majority of workers preferred BL predictions over GT in this example. Then, evaluation boils down to checking if the predicted images describe the input story as adequately as ground truth does. The correspondence between ground truth and predicted images may be caused by the presence of common salient objects/scenes in the images. For example, both GT and BL image sequences in Figure 4.4 visually describe a graduation day scenario. Given that the SIS are not specific with respect to entities, the BL can be considered a correct representation with respect to GT. We propose a visual saliency based similarity metric to evaluate this kind of correctness of a predicted story.

### Text Processing:

We consider a test subset with DII data available. Each image has three captions associated with it. The captions are processed using Stanford core NLP [148] parser to extract noun entities. Some abnormal entities are extracted due to spelling, grammar and typographical errors (e.g., the word ‘advertisement’ had five different spellings). Even though ‘autocorrect’ \(^1\) corrected most of them, the corrections were not always acceptable (e.g. “abike” was changed to “alike” instead of “a bike”). Others like “PyEnchant” \(^2\) required manual verification. To automate the process of correcting thousands of words, we use autocorrect

---

\(^1\)https://github.com/phatpiglet/autocorrect/

\(^2\)https://pythonhosted.org/pyenchant/
She loves the winter and decided to take her new camera out for pictures of the snow. She got a great shot of the trees with clear blue sky in the background. Another beautiful picture of a thin small tree leaning over from the weight of the snow. This was a distance shot against the blue sky. And finally a great shot of a tree with some color peeking out from under the snow.

Figure 4.5: A sample that our proposed method (bottom) gets less votes from AMT users than the ground truth (middle). However the two image sequences look visually similar.

and consider only modifications for noun entities. The extracted entities generally represent objects and scenes present in the images. There is a total 13,000 unique entities over the entire set.

**Visual Processing:**

We train a VGG-19 [144] model on ImageNet for the 20,754 categories [149] and classify the images of the story test set using this network. Out of the 13,000 entities from captions for images in the dataset, only 2815 matched ImageNet categories. About 95% of the images with captions had more than three entities within these 2815. Hence, we did not need to do further processing to match entities to categories in ImageNet. The top-10 most probable categories are chosen as they are mostly interchangeable. These categories were too specific compared to entities extracted from the VIST dataset (e.g., image of a ‘daisy flower’ had ‘flower’ as an entity in the descriptions while the exact type of daisy was the predicted result from ImageNet). Hence we take immediate two hypernyms of the predicted labels using WordNet [150] for each of the 10 categories to make the visual label
list. The union of the visual label list and textual entity list, made up the salient entity set that has both visually and textually grounded entities.

**Evaluation Metric:**

We provide $\text{Recall}@k$ ($k = \{1, 2, 5\}$) for a story in the top-$k$ predictions to have the same salient entities as the ground truth. For each sentence in the story, we retrieve the top ‘k’ images to get ‘k’ visual stories. If at least one of the ‘k’ stories have for each of the images, more than ‘n’ salient entities common with the GT, then it is positive. ‘n’ is experimentally chosen as 10% of the entities of GT as lower values had erroneous results and higher values had poor Recall.

4.4.4 User Study

We perform evaluation with the help of AMT workers. We obtain prediction results from the network without coherence ($\text{NSI}$), network with coherence ($\text{CNSI}$), baseline ($\text{BL}$), and ground truth sequences ($\text{GT}$). Five experiments are performed: 1) $\text{BL}$ vs. $\text{GT}$; 2) $\text{NSI}$ vs. $\text{BL}$; 3) $\text{CNSI}$ vs. $\text{NSI}$; and 4) $\text{CNSI}$ vs. GT and 5) $\text{BL}$ vs $\text{CNSI}$. We can draw a conclusion that $\text{CNSI}$ is the best model among the three approaches if $\text{CNSI}$ is preferred in (3), (4) and (5). Additionally, results of experiments (1) and (2) shows the performance gain of $\text{NSI}$ over $\text{BL}$.

For the AMT experiment, we take 200 random stories from the test set. The test set consists of 5054 stories with 8187 unique images. We reduce this set to only have stories in which all of the images have captions associated with them. This reduces the dataset to 3384 stories with 5055 unique images. Even within this set, there are only 1013 unique images in sequence, even though all of the SIS are unique. Two image sequences corresponding to the same text obtained from $\text{GT, BL, NSI, or CNSI}$ are presented to the user, who is asked to make a binary selection of which visual story best characterizes the text. The order of occurrence of the two representations are randomly shuffled. Each experi-
Graduation day has finally arrived. All the students started filing in. He was thrilled to finally be graduation. Everyone posed for pictures outside the venue. And to cap off the day; we all hung out at the pool.

4.5 Results and Discussion

4.5.1 User Study

Table 4.1 outlines the pairwise preference results from the AMT user study. For the comparison between BL and GT, BL was preferred 106 times, in comparison to GT which was selected 894 times. Thus, the users’ preference of BL over GT was 10.6%. Preference of BL over GT is 3% (6 out of 200), if 3 of the 5 users (majority) preferred the visual story obtained using BL. The user preference for NSI over BL is 67.7%. Similarly, the user
The stands were absolutely packed at this year’s Virginia Tech graduation ceremony. There were literally seas of people almost as if there was a sporting event. The commencement came with speeches from the top students in the graduating class. I will be keeping the program in remembrance of the graduation.

When she goes out she always has a great time. She loves sitting at the bar; Drinking and talking with the bartender. Friends are always there to have fun also. There is always a lot of drinking; talking; and just having a good time!

For the halloween party; Linda dressed all in black. She is chatting with friends she hasn’t seen in a while. Everyone has such great costumes; one lady even came dressed as riding hood. Another guest decided to come as a Pokemon character. Even little kids dressed up for this party.

Figure 4.7: Samples of predicted images for three stories where the image sequences in the last row predicted by our CNSI model wins the most votes from AMT workers.
Table 4.1: The results of pairwise preference test on story visualization of workers reviews via AMT. Comparisons are conducted in the manner of A vs. B. The numbers indicates the percentage of responses that A is a better visualization than B for a given story.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>#Votes</th>
<th>BL</th>
<th>GT</th>
<th>NSI</th>
<th>CNSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL vs. GT</td>
<td></td>
<td></td>
<td></td>
<td>52.5</td>
<td>44.5</td>
</tr>
<tr>
<td>10.6% (106/894)</td>
<td>#Votes</td>
<td></td>
<td></td>
<td>53.3</td>
<td>55.5</td>
</tr>
<tr>
<td>CNSI vs. GT</td>
<td></td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>22.2% (222/778)</td>
<td>#Votes</td>
<td>3</td>
<td>3</td>
<td>20</td>
<td>37</td>
</tr>
<tr>
<td>NSI vs. BL</td>
<td></td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>67.7% (677/323)</td>
<td>#Votes</td>
<td>44</td>
<td>57</td>
<td>54</td>
<td>26</td>
</tr>
<tr>
<td>CNSI vs. NSI</td>
<td></td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>53.3% (533/467)</td>
<td>#Votes</td>
<td>15</td>
<td>41</td>
<td>56</td>
<td>47</td>
</tr>
<tr>
<td>BL vs. CNSI</td>
<td></td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>38.5% (385/615)</td>
<td>#Votes</td>
<td>3</td>
<td>17</td>
<td>40</td>
<td>62</td>
</tr>
</tbody>
</table>

Table 4.2: Preference of algorithm based on maximum voting of 5 workers for 200 samples. To avoid ties, if GT and CNSI get 2 votes each and BL gets one, then both GT and CNSI gets half a point.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>GT</th>
<th>BL</th>
<th>NSI</th>
<th>CNSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Samples</td>
<td>52.5</td>
<td>44.5</td>
<td>47.5</td>
<td>55.5</td>
</tr>
</tbody>
</table>

preference for CNSI over NSI is 53.3%. The results indicate that the proposed model outperforms the baseline and users prefer image sequences that are coherent and consistent.

Figure 4.7 shows example visual stories obtained from GT, BL, NSI, and CNSI, where users preferred the results generated using the proposed model (CNSI). Figure 4.6 shows an example scenario where users preferred BL over CNSI.

From Table 4.2, we can also see that when shown all the four results, users tend to prefer the result by the proposed algorithm more than even GT. Both NSI and CNSI perform reasonably well, but we think CNSI could perform significantly better if dataset was more coherent and/or if contribution of coherence is more dynamically modulated by the network (e.g., through some form of attention).
Last weekend we had a great time at the party. Some people had brought some very elaborate costumes. There were a lot of strange people there too. The children had fun playing together. Everyone gathered in the room for the meeting.

4.5.2 Importance of Consistency and Coherence

In terms of the predicted images, consistency indicates visual similarity between images of a sequence while coherence is interpreted as images having common entities, such as a person or an object. In Figure 4.8, it is clear that images of GT (middle) show higher visual consistency and coherence than our prediction (bottom). However, our predictions were preferred by majority of workers for the visual description of the input story. The fact is that users’ preference is highly related to the alignment between the images and the corresponding sentences apart from visual consistency and coherence. Ideally, consistency and coherence in the output sequence is preferable as shown by the results in Table 4.1 but not always. For example, a set of 5 images that lack visual coherence can still be perceived by a user as forming a story. This is the case for many samples in the VIST dataset. Also, in [de], authors show failure cases that result from giving order to unordered sequence of images and sentences within the same story in the VIST dataset. They observe that failure cases were due to lack of coherence in the dataset itself. This motivates explicit encoding of coherence in the input story, but not constraining the predicted images to have common
objects and scenes.

4.5.3 Visual Saliency Metric

<table>
<thead>
<tr>
<th></th>
<th>Recall @ 1(%)</th>
<th>Recall @ 2(%)</th>
<th>Recall @ 5(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL</td>
<td>5</td>
<td>31</td>
<td>54</td>
</tr>
<tr>
<td>NSI</td>
<td>16</td>
<td>30</td>
<td>43.5</td>
</tr>
<tr>
<td>CNSI</td>
<td><strong>20.5</strong></td>
<td><strong>33.5</strong></td>
<td>50</td>
</tr>
</tbody>
</table>

Table 4.3: Visual Saliency based Recall @1, 2 and 5.

Table 4.3 show the Recall at 1, 2 and 5 for a predicted sequence of images to be visually similar to the images in ground truth. Visual similarity can be explicitly verified in Figure 4.5, where an example story (bottom) predicted by CNSI was in top 1 with respect to GT. However, GT was preferred by majority of workers. Hence, we believe that user study alone or the metric alone would not suffice to measure the performance of the proposed algorithm. Even though there exists some mismatch between the results, we can see a clear pattern with respect to which models perform the best on the dataset. We can also see that, as $k$ in Recall at $k$ increases, performance of BL starts to increase more than that of the proposed model. This might be because, as the number of considered images increase, finding an image with same visual entities as GT become easier while finding images that adhere to the story and are also visually similar to GT becomes more difficult. The values

<table>
<thead>
<tr>
<th></th>
<th>@10(%)</th>
<th>@ 50(%)</th>
<th>@ 100(%)</th>
<th>@ 500(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>3</td>
</tr>
<tr>
<td>NSI</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>CNSI</td>
<td>0</td>
<td>1.5</td>
<td><strong>4.5</strong></td>
<td><strong>24.5</strong></td>
</tr>
</tbody>
</table>

Table 4.4: Visual Saliency based Recall of GT images@10, 50, 100 and 500.

for Recall of retrieving GT images are shown in Table 4.4. we can see that the proposed algorithm performs better though GT images are not retrieved in ranking top 10 or 20. Note that GT images are not the unique visual representation of input story [135] and sometimes stories retrieved by our algorithm are preferred as shown in Table 4.2.
4.6 Limitations and Improvements

We create a novel task, Story Illustration, the automatic illustration of an input natural language story with a sequence of images. The task imitates the human visualization process when comprehending a story thereby combining comprehension with composition. We believe that the proposed task on existing VIST dataset is better suited for story comprehension.

We propose a two stage network as a solution for the problem of visual illustration of natural language stories. Two networks, along with a baseline were evaluated on a comprehensive dataset using both qualitative and quantitative metrics. We observe that the proposed model performs better than the baseline and in a few cases, better than the ground truth itself, as verified by the user study. We observe that evaluation metrics for the storytelling task is ill-defined and hence propose a visual saliency based recall metric as the new measure. Though the metric does not evaluate the correctness of a visual story by itself, when considered along with the user study, it gives useful insights on the results. It is observed from evaluation that this task is non-trivial and more research is necessary to study the relationship between SIS and corresponding images.

Even though the proposed networks perform well, we note that there is a big space for improvement. Particularly, studying the importance of coherence for visualization, developing a more comprehensive and proper dataset and generalizing the approach to a many–to–many mapping between input story and output image sequence. We are looking to explore these paths to define a more robust and well defined solution to the problem of visual storytelling.
CHAPTER 5
MODELLING COHERENCE AND MANY-TO-MANY MAPPING

Story Illustration is the task of illustrating a natural language story with a coherent sequence of images. In chapter 4 we propose a two stage hierarchical neural network that relies on learning one-to-one mappings between text segments and the corresponding images, for story illustration. Though our technique performs well, we observe that neither the dataset nor the model can be generalized to situations where images and text do not have one-to-one correspondence. On the contrary, in storytelling [42], text decoders can generate stories of any length in an open ended manner. This constraint limits the applicability of these techniques to real-world applications such as story comprehension [14], completion or recommendation from scripts [151], real-time visualization and storyboarding [152], food images and recipe modelling [153, 154] where images and text need not exhibit explicit one-to-one correspondence. Moreover, segmenting a coherent textual story into distinct parts is non-trivial [155], causing the associated images to often have overlapping correspondence.

Since our work on CNSI (chapter 4), there has been extensive research on story illustration using VIST dataset. [156] proposed a GAN framework for story generation instead of illustration, evaluated on cartoon dataset [157]. [152] proposed a dense visual semantic matching and context-aware hierarchical story encoding for story illustration on the VIST dataset. Other similar works include [155] that aligns photo streams with text segments of a blog while [135] formulate the problem of sorting jumbled images and captions to form a coherent story on the VIST dataset. [153] propose a variational recurrent network for step wise illustration of cooking recipes.

However, it is shown in [3, 127, 154] that visual coherence in the sequence of images in VIST dataset is highly variant and sometimes non-existent. Consequently, the stories are
too abstract with limited grounding, increasing the ambiguity in details that could go between two consecutive images or time instants. Also, these techniques model relationship between sentences in a story and then map each sentence to one image as constrained by the dataset, restricting its applicability to many-to-many settings. Though some techniques [3] model global textual coherence, absence of visual coherence in the dataset makes the problem of coherent story illustration ill-posed. Moreover, trained models like CNSI section 4.3, NSB [152], etc. are not designed to explicitly capture the relationship between images in the illustration of a story.

5.1 Generalizing Story Illustration to Many-to-Many

In this chapter, we propose the generalized task of Many-to-Many Story Illustration (MMSI) where an input textual story is illustrated with a coherent image sequence (Figure 5.1), by retrieving the images sequentially from a large pool of images. We do not assume any one-to-one correspondence between text and images. Note that this is not the same as the text-to-video retrieval task [37, 38, 105] where videos are an already ordered set of images. In our case, the model has to retrieve the images in a coherent fashion from a large pool of presumably random images, conditioned on the input story. This is a more generalized imitation of the visualization of events in our minds when we read and comprehend a story [139] compared to the one–to–one formulation. We also propose Many-to-Many Story Completion (MMSC) task, where part of the image sequence is provided as additional input with the input text, during inference, and the model retrieves the remaining images forming a coherent sequence. This is similar to the oracle model in [158] or cloze-style tasks described in [159] and [160].

The MMSI task poses important challenges. First, to the best of our knowledge, there are no meaningful datasets that support coherent many-to-many story illustration and existing datasets are one-to-one and lack visual coherence in the stories [3]. Second, text generators [20], in principle, implicitly model coherence using global context from text in
Another dirt clod hits PERSON1_M in the arm. PERSON1_M tries to run along the road, but his braces makes it impossible. He hobbling along as PERSON2_F yells after him. Boy #1 and Boy #2 turn back towards the bikes. The boys ride after PERSON1_M.

Figure 5.1: Example input story and corresponding illustration (ordered row wise starting from top left) from the created dataset. Note the coherence between images in the illustration.

the data. However, coherence within an image sequence for story illustration is conditioned on the input story. The order could be meaningfully interchanged for a slightly different input story making it hard to learn global visual coherence. Also, this makes evaluation of the task non-trivial, requiring subjective human evaluation [3]. Third, various decoding techniques such as temperature sweep [161], beam search [162] and use of special ⟨EOS⟩ token to learn sufficiency, help improve the performance of text decoders but have not been investigated for the proposed setting of image sequence retrieval. Finally, though MMSI task involves retrieval [10] and ordering [163] that are both challenging tasks in itself, it is important to study the overall task for its connection with visualization as a measure of story comprehension [14]. When humans visualize while comprehending stories, the sub-tasks are entangled and influence each other. This resembles a machine translation setting where the image sequence being visualized is an alternate representation of the natural language story.
Problem Formulation

Let \( S = [w_1, w_2, ..., w_m] \) be a natural language story composed of \( m \) words. Though in principle \( w_i \) could be words, phrases, sentences or any other semantically viable unit of text. Similarly, let \( I = [i_1, i_2, ..., i_n] \) be the sequence of images that best describe the input narrative. Note that we do not assume one-to-one mapping between the sequence of semantic units of input text and sequence of images, i.e. \( m \neq n \).

MMSI:

Then the MMSI task is to retrieve \( I \) given \( S \). Formally in our case, \( I = \arg \max_I \Pr(I | S) \).

MMSC:

Equivalently, if \( I_k \) is a non-consecutive sequence of \( k \) images from the set of all images \( I \) for that story, then we can define the MMSC task as \( I_{\sim k} = \arg \max_{I_{\sim k}} \Pr(I_{\sim k} | S, I_k) \), where \( k \ll n \) and \( I_{\sim k} = I - I_k \).

Contributions

We address these challenges with the following contributions:

(1) We develop a novel MMSI dataset by aligning narrative description with images sampled from video clips and propose MMSI and MMSC tasks. While the proposed dataset is based on MPII movie description dataset [133], it enables the investigation of a new class of tasks and the study of visual coherence and many-to-many mapping between sequence of images and text. We also evaluate the applicability of the dataset by conducting a user study showing that stories in the created dataset are comprehensible and coherent.

(2) We propose an end-to-end encoder-decoder architecture that sequentially retrieves image sequences of arbitrary length for an input textual narrative, trained using triplet margin loss [164] with online hard negative mining [165]. We leverage text generation techniques
PERSON1_F stops at a bakery and walks down the street, unashamedly munching on an onion bagel. The fashionable ladies flag down taxicabs and get into chauffeur-driven cars, whilst PERSON1_F makes her own way down a busy city street, as smartly dressed business people dash this way and that. Skyscrapers loom overhead on either side of the street. She arrives at a tall building housing Ellas–Clark Publications and gazes up at its impressive facade.

Figure 5.2: An Example story from the proposed dataset. High visual coherence and visual grounding is visible. The text is detailed yet narrative in nature capturing the storytelling intent.

like greedy decoding and curriculum learning to successfully aid the network in modelling coherence in the retrieved images. To the best of our knowledge, text generation type open–ended decoders have not been investigated for the proposed setting of sequential image retrieval.

(3) We perform comprehensive quantitative analysis and ablation studies on the proposed MMSI dataset, showing better performance than baselines. Further, we evaluate the efficacy of the proposed technique, on an existing one-to-one storytelling dataset, VIST [42]. On VIST, our model achieves performance comparable to more complex baselines that are tailored towards the one-to-one setting showing that the proposed model can generalize well to one-to-one datasets as well.

(4) Given the subjective nature of the task, we qualitatively evaluate the proposed model with a user study. The proposed model attains scores closer to the ground truth in terms of coherence and relevance.
5.2 MMSI Dataset

5.2.1 Necessity

Numerous datasets have been created such as MSCOCO [12], Visual Genome [166], Conceptual Captions [167] each with different characteristic and features, for one-to-one image-text alignment tasks like captioning, segmentation or recognition. Story Illustration is a much more complex task with limited existing literature. It is imperative to model stories [14] for the development of general machine comprehension systems and to achieve true visual and textual reasoning. There are various factors that govern story illustration including context, coherence, causality between events and creativity, to name a few. To the best of our knowledge, there are limited datasets like VIST [42] that supports visual storytelling albeit with limitations as explained above. The proposed MMSI dataset does not by itself eliminate the need for existing datasets, but aim to complement those like VIST by providing support to model different factors such as visual coherence or many-to-many mapping. Specifically, the proposed model for sequential open-ended illustration on the proposed dataset can be applied to VIST as shown in Table 5.4, while models trained explicitly using the one-to-one constraint in VIST [3] are not directly applicable to the proposed dataset. This also shows the generalization ability of the proposed dataset. This dataset is similar to the RecipeQA dataset [163] but for more general form stories that are basically events in movies.

5.2.2 Description

We create the MMSI dataset from the MPII movie description dataset introduced in [45]. MPII dataset consists of textual narratives aligned with video clips. To achieve many-to-many correspondence, we combine five consecutive narrative descriptions and their corresponding clips as described in the Large Scale Movie Description Challenge (LSMDC)
Then, we sample frames from each of these five clips by extracting the key frames\(^1\) from the clips to give an arbitrary number of *important* frames. This gives us a narrative made of minimum five sentences and an arbitrary number of images. To avoid long sequences, we limit the number of key frames per clip to a maximum of five via uniform sampling. The average number of words per story is 53. Minimum and maximum number of words per story are 18 and 179 respectively. The vocabulary size for our models is 25,024. We do not truncate the vocabulary as the model does not involve decoding of text. The number of images per story varies between 20 and 25. The distribution in percentage, of the number of images per story is 1.7%, 4.9%, 11.7%, 21.7%, 30.1%, 29.9% for 20, 21, 22, 23, 24 and 25 respectively. The clips and stories were obtained from a total of 184 movies. There are 101,079 and 10,053 clips in training and test splits totally. About 60% of the training clips have more than 5 frames. Before sampling, there are a total of 821,846 frames in the training set which is reduced to 452,423 frames after uniformly sampling 5 key frames throughout the video clips with more than 5 key frames. The maximum number of images per story is then 25. We do this for the entire publicly available data in the LSMDC challenge [133] dataset to get 19,151 training, 1,000 validation and 1,477 test stories. In total there are 452,423 images in the train, 23,954 in the validation and 48,133 in the test set.

The proposed dataset can be obtained by using our code on the publicly available MPII movie description and LSMDC datasets. We refer to the original articles [45, 133] for details on the original videos and accompanying text. Detailed instructions to create the dataset are given in the code accompanying supplementary materials. The proposed dataset has approximately five times more images per story compared to VIST but spanning a much shorter time period. There is considerably more coherence in the visual stories as shown in Figure 5.2 and Figure 5.3. The corresponding text also has relatively more details than the abstract stories in VIST. More importantly, though the text is detailed, it still

\(^1\)https://ffmpeg.org/about.html
More cannonballs explode the PERSON1 lines, claiming more victims. PERSON2 fires his gun blind, the smoke obscuring his vision. PERSON3 tosses a bag at PERSON4 and hares away across the fields, his physical and mental strength restored. The two of them dive into thick woodland, moving behind enemy lined, with a view of German soldiers and their tank ahead of them. They crouch down, as half a dozen Germans come past and set up a gun position in front of them.

PERSON1 considers this a moment, then attacks, and if he moved quickly last time, this time he is blinding and as PERSON2 slips down to avoid charge, PERSON1 moves right with him, only instead of twisting free and jumping to his feet, this time PERSON2 jumps for PERSON1’s back and in a moment he is riding him, and his arms have PERSON1’s throat, locked across PERSON1’s windpipe, one in front, one behind. PERSON1, as he charges toward a huge rock, spins his giant body so that entire weight of the charge is taken by PERSON2. And the power of the charge is terrible, the pain enormous, but he clings to his grip at PERSON1’s windpipe.

Figure 5.3: More example stories from the proposed MMSI dataset. All the stories show high visual and textual coherence.
PERSON1_M and the girl seem blissfully unaware of them. Turns again and they row past PERSON1_M and the girl, PERSON2_M again clicking off several fast shots. Move along and the red tiled roof and down to a lower level of the roof where PERSON2_M’s feet are hooked over the apex of the roof and PERSON2_M himself is stretched face downward on the tiles, pointing himself and his camera to a veranda below him where the girl and PERSON1_M are eating. PERSON2_M is clicking off more shots when the tiles his feet are hooked over come loose. PERSON2_M begins a slow slide down the tile to the edge of the roof anf possibly over it to a three-story drop.

Figure 5.4: An example story with intricate details (2nd and 3rd sentences) in the text not explicitly captured in the visual.

follows a narrative style rather than being descriptive or factual about objects and scenes in the image like in [12]. There is intuitively more visual grounding and lesser chances of ambiguous text filling the gaps between images in a story. This is particularly desirable in the preliminary stages of story illustration and a well trained model on this data could then meaningfully be extended to be ‘creative’ in filling the gaps in more abstract stories. We do observe certain stories with intricate details not captured in the visual stories as shown in Figure 5.4.

5.2.3 Dataset Evaluation

We perform a user study to assess the applicability of the proposed task and the dataset. The input story, ground truth image sequence and a shuffled version of the ground truth image sequence of 100 random stories from the proposed and VIST datasets are shown to Amazon
Mechanical Turk (AMT) workers. They are asked to identify the ground truth sequence. Image sequences of fifty stories are randomly shuffled while those of the remaining are first divided into five equal parts and then two of the parts are interchanged, to make the task harder. Each story was shown to five workers and we consider an example as success if three out of the five workers correctly identify the ground truth sequence from the shuffled sequence.

We observe that 81% of the stories in the proposed dataset were identified successfully, indicating that the stories and their alignment to image sequences are comprehensible in the proposed MMSI dataset. Interestingly, 82% of randomly shuffled stories and 80% of the carefully shuffled stories were a success showing negligible difference between the methods used to shuffle the sequences.

VIST dataset on the other hand got a success rate of 85%, an increase of 4% over the proposed dataset. In VIST, there are only five images per story with each image illustrating one sentence. We believe that its easier to identify ground truth sequence from a shuffled sequence by leveraging the one-to-one correspondence, despite the low visual coherence between images. We also did an experiment where image-text pairs are interchanged rather than just the images. Only 62% of the stories were correctly identified showing that stories in VIST lack coherence making it unsuitable for coherent story comprehension. In the proposed MMSI dataset however, each story has a minimum of 20 images, four times as much as that of VIST and identifying a coherent sequence from a shuffled sequence requires comprehending the narrative and illustration as a whole. Our study shows that the proposed MMSI task and dataset is a meaningful and challenging benchmark for studying story illustration.

5.3 Machine Translation for MMSI

We propose an end-to-end neural model based on sequence to sequence architecture as shown in Figure 5.5. The proposed model consists of a text encoder $f_e(\cdot)$, text-to-image
Another dirt clod hits PERSON1_M in the arm. PERSON1_M tries to run along the road, but his braces makes it impossible. He hobbles along as PERSON2_F yells after him. Boy 1 and Boy 2 turn back towards the bikes. The boys ride after PERSON1_M.

Figure 5.5: Proposed Encoder-Decoder training framework for story illustration. (A) is the sentence encoder (B) story encoder (C) text-to-image decoder and (D) is the trainable part of image encoder. During inference, image closest to $\hat{I}_t$ (generated by text-to-image decoder) is retrieved and then given as input for retrieving subsequent images.

decoder $f_d(\cdot)$ and an image encoder $g(\cdot)$ all trained jointly using the triplet-margin loss function with online hard negative mining strategy.

5.3.1 Text Encoder

The text encoder (Figure 5.5) is a hierarchical sentence-story GRU network similar to [152, 3]. The sentence encoder is a two layered Bidirectional GRU that sequentially encodes the word embedding of each word $w_i$ to get a sentence vector independently for each of the sentences. The sentence vectors for each sentence are then sequentially encoded by another two layered bidirectional GRU called Story Encoder. We get each sentence’s contextual hidden state $z_i \in \mathbb{R}^d$ described as $f_e(S_i) = z_i$ while the final hidden state $z \in \mathbb{R}^d$ represents the overall context of the entire input story. Outputs of the BiGRUs are first concatenated along both directions and then linearly encoded to the required dimensions. Sentence level outputs and overall context have different linear layers to model global and local context independently.
5.3.2 Image Encoder

For the image encoder (Figure 5.5), we first extract 2048–D features from last layer of ResNet-152 [168] model pretrained on ImageNet [168]. These features are then passed through two trainable linear layers with ReLU (Rectified Linear Unit) [169] activation function and a batchnorm layer [batchnorm] in between them to get $g(i_k) = I_k \in \mathbb{R}^d$ for each image in the story.

5.3.3 Text-to-Image Decoder

The text-to-image decoder (Figure 5.5) $f_d(\cdot)$ retrieves the sequence of images $I$ conditioned on encoder outputs $z_i$, context vector $z$ and previously retrieved image $I_t$. In order to indicate the end of a sequence, we use $\langle$EOS$\rangle$ whose input representation $i_{eos}$ is a 2048–d vector of ones ($\rightarrow 1$). The corresponding embedding $g(i_{eos}) = I_{eos}$ is learned while training the image encoder. Similarly $\langle$SOS$\rangle$ has input $i_{sos}$ as a $\rightarrow 0$ vector to indicate start of sequence for the decoder. We use $\langle$PAD$\rangle$ with input $i_{pad}$ as a $\rightarrow 0$ vector for padding. A ground truth image sequence of length $n_{max} – 2$ excluding $\langle$SOS$\rangle$ and $\langle$EOS$\rangle$ is then represented as $\{i_{sos}, i_1, i_2, ..., i_{n_{max}-2}, i_{eos}, i_{pad}, i_{pad}\}$ where $n_{max}$ is the maximum number of images per story.

The core of the decoder is a two layer GRU network with cross-attention [170] over the text-encoder outputs. The first token to the decoder is always $g(i_{sos}) = I_{sos}$ while its hidden state is initialized using the final hidden state, $z$ of the text encoder. At each time step, the current input $I_t$ is passed to the decoder GRU to get $h_t$. This step learns the necessary visual context based on the current image. We then use $h_t$ to apply attention over $z_i$ to measure the importance of each encoded word (or sentence) $z_i$ as $c_t = \sum_{i=1}^{m} a_{i,t} \cdot z_i$ where $a_{i,t} = \exp(e_{i,t}) / \sum_{j=1}^{m} e_{j,t}$ are the attention weights while $e_{i,t} = W_a[h_t; z_i] + b_a$.

The context $c_t$ is mapped to $\mathbb{R}^d$ via $W_b[c_t; h_z] + b_b$. We also tried the scaled dot product multi-head attention as proposed in [104] but observed better performance with the proposed linear attention. Output is then added with the previous image’s feature to ensure
coherence and passed through a linear layer to get $f_d(I_{sos}) = \hat{I}_1$. During inference, we retrieve the image in test set with the closest image embedding to $\hat{I}_1$. The retrieved image vector $I_1$ is then given as input for the next time step. This process follows for all $I_t$ until the retrieved vector is $I_{eos}$ or when the number of images retrieved is $n_{max}$.

5.3.4 Loss Function

The entire network is optimized in an end-to-end fashion using *triplet margin loss* [164] with online hard negative mining [165]. If $d(x, y) = \|y - x\|^2$ characterizes the euclidean distance between two vectors and $\text{trip}(a, p, n) = \alpha + d(a, p) - d(a, n)$ is the triplet loss, then the loss function is given in Equation 7.1.

$$c = \sum_{k=1}^{bs} \sum_{t=1}^{t_k} \left( \max\{0, \text{trip}(I_{k,t}, \hat{I}_{k,t}, I'_{k,t})\} \right) + \max\{0, \text{trip}(\hat{I}_{k,t}, I_{k,t}, \hat{I}'_{k,t})\} \right)$$

(5.1)

where $\hat{I}_{k,t}$ and $I_{k,t}$ are corresponding predicted and ground truth images respectively. Following online hard negative mining strategy [165], $I'_{k,t}$ is an image sampled from all predicted and ground truth images that do not correspond to the current ground truth image $I_{k,t}$ but has the smallest distance to ground truth from the current batch. This minimizes the distance between a positive pair of images while ensuring that the closest non-corresponding image in the batch is farther by a margin $\alpha$.

5.3.5 Sequential Image Retrieval Techniques

Inspired by Text Generation [171, 172] techniques, that use sophisticated decoding strategies during inference [172] to improve the performance of a trained model we consider three different decoding strategies for sequential image retrieval for story illustration.

Direct.

We directly give the predicted vector $\hat{I}_t$ generated by the decoder, as input to the next time step.
Greedy.

Similar to greedy decoding [172] in text generation, during inference, we retrieve the image with embedding closest to the predicted vector $\hat{I}_t$ at each time step $t$ from the set of image embeddings $\{I_1, I_2, ..., I_N\}$ according to the violation described in Equation 7.1. Then the retrieved image’s vector $I_k$ becomes the input for the next time step.

Beam Search.

Beam search [162] has been shown to give significant improvements over naïve greedy decoding [172]. In our case, instead of choosing the top-$k$ tokens with maximum probability at each time step $t$, we choose the top-$k$ image vectors with the lowest distance under the violation described in Equation 7.1 where $k$ is the beam size. This continues until $I_{eos}$ is retrieved as the vector or until $n_{max}$ is reached for all the sequences. The path with the lowest overall score is the final sequence.

5.3.6 Training Methodologies

Exposure bias [173] is a very common problem when ground truth inputs are used during training, while during inference, the model decodes using its own predictions. To address this, strategies like curriculum learning and scheduled sampling have been proposed. We evaluate the following training methodologies to overcome exposure bias [173].

Teacher Forcing.

In this method, we give the ground truth image vectors $I_t$ as input to the decoder for predicting the next image during training.

Curriculum Learning.

We sample the ground truth image with probability $1-p$ and the predicted vector $\hat{I}_t$ with probability $p$ at each time step. We start with $p = 0.05$ and increase it until 0.45 with a step size of 0.05 every 5 epochs.
5.4 Experimental Setup

We evaluate the model on two tasks. For the first task (MMSI), given the entire text story and $I_{sos}$ the model retrieves a sequence of images $[i_1, i_2, \ldots, i_n]$ to illustrate the story. In the second task MMSC, we evaluate the model trained for Story Illustration when ‘k’ images from the ground truth illustration are given by treating them as the model’s predictions and the model is required to retrieve the remaining $n_{max}$-k images. Notation $t_{1,5,10}$ indicates that the first, fifth and tenth images in the story are input seeds.

5.4.1 Training Details

The hidden dimensions of the encoder and decoder are 512 and the image encoder and text-to-image decoder’s outputs have a dimension of 1024. Maximum number of words considered per story is 200 for Bi-GRU encoder, while the maximum sentences per story is 10 and maximum words per sentence is 20 for the Hierarchical Sentence Story Encoder. The maximum number of images that is part of a single story is 25 considering 5 frames per clip that is used to create the sequence of images. We use the Adam optimizer [174], initialized with a learning rate of $1e^{-3}$ with a linear decay by a factor of 0.5 every 5 epochs. We train all the models for 100 epochs on the training set with linear learning rate decay for all the modules. The epoch with the highest Mean Reciprocal Rank (MRR) on validation data is chosen as the best epoch and evaluated on the test set. During inference, to calculate the metrics, we retrieve top k images each time step. For beam decoding, metrics are calculated using all the k retrieved sequences. Median Rank with retrieval range of 1000 is calculated by sampling 1000 random images from test set along with ground truth, averaged over five runs. We choose the margin $\alpha$ in the objective as 0.5 from $\{0.1, 0.2, 0.3, 0.4, 0.5, \ldots, 2.0\}$. Learning rate of $1e^{-3}$ was empirically chosen from $\{1e^{-4}, 5e^{-4}, 1e^{-3}, 5e^{-3}, 1e^{-2}, 5e^{-2}\}$, hidden dimension of 512 for the networks from $\{128, 256, 512, 768, 1024\}$ and batch size from $\{8, 16, 24, 32, 48, 64\}$. We tried to limit the vocab size to 10 000 and 20 000
but observed a marginal decrease in the performance and hence kept the original vocabulary of 25024 as it is. We also tried various methods to initialize the word vectors such as using glove-300 [175] and training a word2vec model using gensim [176] but noticed no increase in the performance. The total training time over the entire dataset for the proposed model per epoch is 10 minutes for a batch size of 32 on a single Nvidia Quadro K6000 GPU. During inference, it takes about 2 minutes to retrieve top10 images for all samples in the test set with greedy decoding and about 10 minutes for beam size of 2 and 55 minutes for beam size of 10 in the same GPU.

5.4.2 Evaluation Metrics

For evaluating text sentences of different length, there are many language evaluation metrics [177]. However, they cannot directly be used for evaluating image sequences of varying lengths. We use rank based metrics: Recall@k (R@k), median Rank (MedR) and Mean Reciprocal Rank (MRR). These measure the ability of the model to retrieve the ground truth image as the correct image per story per instant. We ensure that the metrics are calculated until \( \langle \text{EOS} \rangle \) in ground truth sequence. Additionally, to check if the model is predicting sequences of same length as ground truth, we calculate the Mean Absolute Error between the lengths of predicted and ground truth sequences over the entire data. We also employ another metric called Story Recall@k inspired by [153]. While Recall@k measures the ordering of the images, story recall relaxes that constraint and checks if a predicted image at a given time step is the same as any of the ground truth images within the same story.

5.4.3 Baselines and Comparisons

Ablation Study

To the best of our knowledge this is the first model to perform sequential image retrieval for MMSI task. Though recently developed architectures based on multimodal transformers
<table>
<thead>
<tr>
<th>Model</th>
<th>Hierarchical Encoder</th>
<th>Attention</th>
<th>Curriculum Learning</th>
<th>Decoding Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>Greedy</td>
</tr>
<tr>
<td>$P_{Teacher}$</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>Greedy</td>
</tr>
<tr>
<td>$P_{NoAttn}$</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>Greedy</td>
</tr>
<tr>
<td>$P_{Direct}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Direct</td>
</tr>
<tr>
<td>$P_{BeamK}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Beam</td>
</tr>
<tr>
<td>Proposed</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Greedy</td>
</tr>
</tbody>
</table>

Table 5.1: Description of the models used for comparison. The prefix ‘P’ indicates that the proposed model is used except for the indicated change. In ‘Baseline’ model, Decoder attention is removed and Hierarchical Sentence Story Encoder is replaced with a Bi-GRU. $P_{Teacher}$ model is trained with Teacher Forcing.

[11, 62, 105] might help improve the performance, they encode textual and visual information onto a joint space rather than translating one to another. This is applicable for text-to-video retrieval tasks where video is a single already ordered entity but, not for the proposed MMSI task that requires sequential retrieval of images using a decoder. We leave explorations of such architectures for future work as the proposed setup of sequential illustration similar to translation is novel, generic and adaptable. We perform comprehensive ablation study with respect to architectural design choices, training methodologies and decoding strategies to motivate the final model as described in Table 7.3. We also tried a pretrained BERT [19] as the sentence encoder but did not observe significant improvements.

**Evaluation on VIST Dataset**

Since there are no existing baselines that would directly be applicable for the proposed dataset, we also evaluate the performance of the proposed model on VIST [42] dataset and compare with previous work. We train the proposed model on VIST data splits for sequential many-to-many illustration in contrast to previous work that optimize directly on a one-to-one objective. We compare the results with CNSI [3] and NSB (Neural Story-Board) [152]. We provide Recall@$k$ with $k \in \{10, 50, 100\}$, as given in previous works.
(Table 5.4).

5.5 Results and Discussion

5.5.1 MMSI

Illustration performance of the model shown in Table 5.2, is relatively low compared to its completion counterpart. This is understandable given the complexity of the dataset [63] and the problem formulation. We do observe however, that proposed training and decoding strategies improve the baselines. The model still suffers from error propagation wherein the first retrieved image decides the fate of the entire sequence. Consequently performance for Proposed-\(t_1\) is significantly better by giving just one seed image compared to not giving any.

5.5.2 MMSC

It can be seen from Table 5.3 that the story completion performance of the Proposed model is considerably higher than Baseline. We observe an average increase of 4% in Recall and 5% in StoryRecall by \(t_{1,10,15}\). The higher StoryRecall values indicate model’s ability to get relevant images while the lower recall value indicates the difficulty in ordering the retrieved images. The performance of \(t_{1,10,15}\) is better than \(t_{1,2,3,4}\) though the number of images shown to the decoder is higher in the latter case. This is because, in most stories, 10th and 15th images correspond to different scenes whereas the first four images correspond to the same scene or context.

5.5.3 User Study

We emphasize that Recall and Rank based metrics are harsher and unreliable for story illustration, a subjective and abstract task with multiple possible illustrations for a given input story. Hence, we perform a comprehensive user evaluation of the proposed model. We
<table>
<thead>
<tr>
<th>Model</th>
<th>R@1 ↑</th>
<th>R@5 ↑</th>
<th>R@10 ↑</th>
<th>SR@1 ↑</th>
<th>SR@5 ↑</th>
<th>SR@10 ↑</th>
<th>MedR ↓</th>
<th>MRR ↑</th>
<th>MAE ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>1.4</td>
<td>1.7</td>
<td>1.9</td>
<td>2.3</td>
<td>3.2</td>
<td>3.6</td>
<td>16142 (478)</td>
<td>0.016</td>
<td>0.98</td>
</tr>
<tr>
<td>P_NoAttn</td>
<td>1.3</td>
<td>1.7</td>
<td>1.9</td>
<td>2.8</td>
<td>3.1</td>
<td>3.4</td>
<td>10523 (454)</td>
<td>0.016</td>
<td>1.40</td>
</tr>
<tr>
<td>P_Direct</td>
<td>1.2</td>
<td>1.6</td>
<td>1.8</td>
<td>1.4</td>
<td>2.9</td>
<td>3.5</td>
<td>16605 (466)</td>
<td>0.014</td>
<td>0.98</td>
</tr>
<tr>
<td>P_Beam0</td>
<td>0.0</td>
<td>1.6</td>
<td>1.8</td>
<td>1.7</td>
<td>1.9</td>
<td>2.0</td>
<td>12180 (430)</td>
<td>0.008</td>
<td>1.14</td>
</tr>
<tr>
<td>P_Teacher</td>
<td>1.3</td>
<td>1.4</td>
<td>1.4</td>
<td>1.6</td>
<td>1.8</td>
<td>2.0</td>
<td>16726 (476)</td>
<td>0.014</td>
<td>1.18</td>
</tr>
<tr>
<td>Baseline</td>
<td>1.3</td>
<td>1.6</td>
<td>1.8</td>
<td>2.6</td>
<td>3.1</td>
<td>3.6</td>
<td>10967 (470)</td>
<td>0.016</td>
<td>1.32</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Completion (MMSC)</th>
<th>Proposed-\emph{t}_1</th>
<th>P_NoAttn-\emph{t}_1</th>
<th>P_Direct-\emph{t}_1</th>
<th>P_Beam0-\emph{t}_1</th>
<th>P_Teacher-\emph{t}_1</th>
<th>Baseline-\emph{t}_1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5.1</td>
<td>10.8</td>
<td>14.3</td>
<td>20.3</td>
<td>24.5</td>
<td>29.5</td>
</tr>
<tr>
<td></td>
<td>5.5</td>
<td>10.7</td>
<td>13.4</td>
<td>20.4</td>
<td>23.7</td>
<td>27.9</td>
</tr>
<tr>
<td></td>
<td>4.3</td>
<td>8.5</td>
<td>10.9</td>
<td>14.2</td>
<td>16.7</td>
<td>19.0</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>10.0</td>
<td>13.8</td>
<td>17.1</td>
<td>17.4</td>
<td>17.7</td>
</tr>
<tr>
<td></td>
<td>4.0</td>
<td>9.1</td>
<td>11.2</td>
<td>17.2</td>
<td>22.3</td>
<td>25.1</td>
</tr>
<tr>
<td></td>
<td>3.8</td>
<td>8.1</td>
<td>10.6</td>
<td>16.7</td>
<td>20.8</td>
<td>24.0</td>
</tr>
</tbody>
</table>

Table 5.2: Model evaluation on Story-Illustration and Story-Completion (indicated by the suffix \emph{t}_1 meaning first image is given as seed during inference and the model predicts the remaining) tasks. Med-Rank is calculated over the entire 35 164 images and the number in brackets indicate Med-Rank with a retrieval range [10] of 1000.
randomly select 120 stories from the test set and ask human workers in Amazon Mechanical Turk (AMT), to evaluate the retrieved image sequences of ground truth, Proposed and Proposed-$t_1$ models separately, based on two criteria viz., relevance and coherence, similar to previous work [152, 156]. Relevance measures the alignment between the context of the input story and the retrieved images whereas Coherence measures the continuity within the retrieved images. Three workers evaluate each story and the scores are out of 4 with 0 indicating totally incoherent (irrelevant) to 4 indicating highly coherent (relevant) to measure coherence (relevance). The average scores for coherence and relevance is 3.0 and 2.95 respectively for ground truth indicating high coherence and relevance as expected. The Proposed-$t_1$ was rated 2.87 and 2.77, very close to ground truth with just one seed image. Interestingly, the Proposed model got 2.24 and 2.25 for coherence and relevance which is only negligibly lesser than the completion model. These results are in contrast to the quantitative metrics because of the nature of the task and insufficiency of the automatic metrics. The overall narrative still shows that the proposed model with a simple sequential decoder is able to retrieve coherent and relevant sequences for an input story, despite the non-trivial nature of the task and complex dataset.
A little while later, PERSON1_F has managed to remove most of the dirt and blood from her face. The troll sits across from her and she watches him curiously. PERSON2_M looks down, then gets up in a huff.

Figure 5.6: An example input story, ground truth and the predicted sequences of images by the proposed model with $t_1$ followed by $t_{1,10,15}$. BLUE border indicates seed images, GREEN border indicates exact match to ground truth while YELLOW border indicates the presence of the predicted image somewhere in the ground truth story. No border means the predicted image is not part of the ground truth. BLACK image is the $\langle EOS \rangle$. Recall@1 for $t_1$ is calculated as $\# \text{GREEN} / (\# \text{GT} - \# \text{BLUE}) = 37.5\%$. Similarly StoryRecall@1 is calculated as $(\# \text{GREEN} + \# \text{YELLOW}) / (\# \text{GT} - \# \text{BLUE}) = 50.0\%$. The Recall@1 and StoryRecall@1 for $t_{1,10,15}$ are $31.8\%$ and $54.5\%$ respectively. MAE for this example is 0.0 for both the predictions.
And large flakes of fresh white snow drift down outside the repaired windows of the room. The bed is neatly made. A blue bedspread with white polka dots smoothed across it. White pillow plumped up against the barred metal headboard. The door to the on suite bathroom is ajar.

Figure 5.7: Another example result. The Recall@1 and StoryRecall@1 for $t_1$ and are 54.5% and 100% while for $t_{1,10,15}$ are 54.5% and 95.4% respectively. MAE for this example is 0.0 for both the predictions.
PERSON1_M lands on his back in the snow as his fellow birders rush past him. He peers through his binoculars like the others. A small flock of birds takes flight but one lingers behind. It languidly flutters toward a ridge where PERSON2_M photographs it. The others lower their binoculars.

Figure 5.8: Another example result. The Recall@1 and StoryRecall@1 for $t_1$ and are 4.5% and 27.2% while for $t_{1,10,15}$ are 10% and 30% respectively. MAE for this example is 3.0 for $t_1$ as the model predicted 26th image as $\langle EOS \rangle$ instead of the 23rd image as in ground truth, while it is 2.0 for $t_{1,10,15}$. 
We provide more examples in Figure 5.6, Figure 5.7 and Figure 5.8. All predicted sequences are shown until the ground truth \( \langle EOS \rangle \). The MAE scores comparing the length of the sequences are given in the captions. For example, in Figure 5.8, \( \langle EOS \rangle \) in ground truth is the 23rd image while our model with \( t_1 \) predicts 26th image as \( \langle EOS \rangle \). This gives the model an MAE of 3.0 and only 23 predicted images are shown. The input narrative is shown first followed by ground truth and then predictions.

We observe that the model generally adheres to the input story very well. In Figure 5.6, the model predicts image sequence of same length as ground truth. Moreover, we can see that images 14–19 in \( t_1 \) predictions are not the ground truth images but sufficiently illustrate the last sentence in the story. This is because the text mentions Person2_M with no explicit connection to ‘the troll’. Qualitatively, one could argue that this is a sufficiently good illustration of the input story since the images do show a ‘male that is looking down’ in a similar forest background. We can see similar behaviour in other examples (Figure 5.8) as well with much lower Recall values reflecting the observation in user studies. Similarly in the predictions for \( t_{1,10,15} \) model, we see that images 16–21 are extremely similar to the ground truth images showing the ‘troll’ with similar expressions. However, these are from different scenes (possibly successive) in the movie and as they do not directly correspond to the ground truth, the recall values are affected.

It can be seen clearly how the retrieved images adhere to the context provided by the input story. Story completion with \( t_{1,10,15} \) performs much better than with \( t_1 \) as expected. A strong evaluation based on contextual similarity rather than absolute recall values and modified beam search with look back and correct strategy could alleviate this problem. Nevertheless, a detailed study and understanding of the nuances of image and text sequences of arbitrary length and varying levels of coherence is necessary.
<table>
<thead>
<tr>
<th>Model</th>
<th>R@1</th>
<th>R@10</th>
<th>SR@1</th>
<th>SR@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-$t_{1,2,3,4}$</td>
<td>4.5</td>
<td>11.3</td>
<td>17.3</td>
<td>25.5</td>
</tr>
<tr>
<td>Baseline-$t_{1,10,15}$</td>
<td>6.4</td>
<td>19.3</td>
<td>32.1</td>
<td>47.6</td>
</tr>
<tr>
<td>Proposed-$t_{1,2,3,4}$</td>
<td>7.3</td>
<td>18.3</td>
<td>22.7</td>
<td>33.5</td>
</tr>
<tr>
<td>Proposed-$t_{1,10,15}$</td>
<td>8.8</td>
<td>23.4</td>
<td>37.9</td>
<td>52.9</td>
</tr>
</tbody>
</table>

Table 5.3: Baseline and Proposed models evaluated on Story-Completion task. $t_i$ indicates $i^{th}$ ground truth image is provided. Only subset of metrics are provided owing to space constraints.

5.5.5 Training and Decoding Techniques

Scheduled sampling clearly improves the performance of the model (Refer Proposed and $P_\text{NoAttn}$ vs $P_\text{Teacher}$ in Table 5.2) removing exposure bias. Greedy and Beam decoding clearly outperform direct decoding giving an average of 3% increase over all Recall metrics. Interestingly, beam search under-performs when compared to greedy. Beam search is desirable in text generation when there can be multiple coherent and grammatically correct sentences for a ground truth. In our case, the image sequence is strictly conditioned on the input story. Though one could relax this constraint in the hopes of modelling creativity, that would make evaluation much harder.

5.5.6 Generalization to VIST

We can observe from Table 5.4 that the proposed model’s performance on VIST dataset is comparable to NSB [152] and is better than CNSI [3]. Moreover, the model architecture in NSB is heavily tailored and tuned towards the one-to-one data, with explicit coherence modelling. The models map each sentence in a story to corresponding image making the problem simpler as it need not learn sufficiency (the number of images required to illustrate) or visual coherence to perform well. Also, given the one-to-one nature of the model, it is not directly applicable to the proposed dataset. In contrast, our method is generic and the sequential text-to-image decoder that explicitly conditions future predictions on previously retrieved images naturally leads to coherent sequences. More importantly, the performance
Table 5.4: Models trained and tested on VIST dataset. $t_1$ and $t_{1,2,3}$ indicate story completion experiments.

<table>
<thead>
<tr>
<th>Model</th>
<th>R@10 ↑</th>
<th>R@50 ↑</th>
<th>R@100 ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMSI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>3.4</td>
<td>11.2</td>
<td>17.7</td>
</tr>
<tr>
<td>[152]</td>
<td>13.65</td>
<td>33.91</td>
<td>45.53</td>
</tr>
<tr>
<td>[3]</td>
<td>0</td>
<td>1.5</td>
<td>4.5</td>
</tr>
<tr>
<td>MMSC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed-$t_1$</td>
<td>3.42</td>
<td>11.3</td>
<td>17.8</td>
</tr>
<tr>
<td>Proposed-$t_{1,2,3}$</td>
<td>4.56</td>
<td>13.21</td>
<td>19.8</td>
</tr>
</tbody>
</table>

of the proposed model in completing (MMSC) the story given the first (Proposed-$t_1$) or even the first three (Proposed-$t_{1,2,3}$) ground truth images in VIST dataset, doesn’t improve the performance over not providing any seed images, as much as it did for the proposed MMSI dataset. Note that there are only five images per story in VIST. This behavior can be attributed to the highly variant or non-existent visual coherence and one-to-one mapping in the VIST dataset as also pointed out in subsection 5.2.3.

5.6 Limitations and Improvements

In this chapter, we propose the task of generalized many-to-many story illustration, aimed at understanding aligned text–image sequences of arbitrary lengths. We highlight important limitations with existing datasets like VIST and corresponding techniques such as one–to–one correspondence and lack of coherence. To overcome these limitations, a new dataset with coherent stories obtained from videos and aligned narrative is introduced. Also, we adapt machine translation like neural models to propose a generic neural encoder-decoder framework to sequentially retrieve image sequences of arbitrary length aided by training and decoding strategies leveraged from text generation literature. Furthermore, we evaluate the efficacy of the model and the importance of the proposed dataset both quantitatively and qualitatively.

The proposed model though is simple and effective, has some limitations. Firstly, the
composition of the visualization or illustration is formulated as retrieving a sequence of images from a large dataset. However, generative tasks are more natural, practical and closer to how humans visualize while comprehending a story. An important future direction is to formulate the illustration as a generative task where models can compose from scratch, visualizations for an input story. The current model also suffers from error propagation and beam search in its current form doesn’t help. A look-back and revise strategy for beam search that ‘modifies’ the previously retrieved image based on the path being taken could help. We see that in the \[\text{MMSC}\] oracle setting, if seed images correspond to different contexts or scenes, it becomes easier to retrieve the rest. In future, we plan to explore hierarchical decoder architectures for this. Explicitly modelling the order between images within a story will help reduce the disparity between StoryRecall and Recall values. We also plan to extend this dataset to introduce visual stories with varying levels of coherence by sampling frames (possibly using scene detection\footnote{https://pyscenedetect.readthedocs.io/en/latest/}) techniques \([178]\)) with varying frequency. This will help learn a model with knowledge of abstraction. We believe this opens up interesting new paths for generic understanding of text and visual sequences.
“Examples are the best precept” – Aesop, The Two Crabs

Storytelling is a creative cognitive process [6]. Though the task of Story Illustration (chapter 4 and chapter 5) is a reasonable measure of story comprehension, associated datasets like VIST [42, 137] lack coherence and diversity. Story Illustration as a task is formulated as the retrieval of a sequence of images from a large dataset. Human visualization process is better modelled as generative with the ability to compose visualizations using abstract concepts. Cognitive and Neuroscience literature [109, 5, 110] suggests that visual perception is an abstract and temporal process. Current text–to–image generation systems [35, 179] are far from achieving true compositional ability to create story illustrations from scratch and are difficult to train. Moreover, comprehending stories made of real world images require the modelling of all of world knowledge and it is impractical.

In this chapter, we propose AESOP: a novel dataset that captures the creative process associated with visual storytelling. An example story from our dataset is shown in Figure 6.1. To ensure stories are diverse and creative, we ask workers to create both the visual and textual parts of the story simultaneously from scratch. Inspired by [76, 75, 74], our dataset employs abstract visual scenes, with a broad set of choices for objects and attributes needed for visual storytelling. Our dataset emulates the visual environment of comics/cartoons by constraining the number of objects and scenes to a set of pre-defined clip–arts. Using AESOP, we propose foundational storytelling tasks that are generative variants of story cloze tests, to better measure the creative and causal reasoning ability required for visual storytelling. We further develop a generalized story completion framework that models stories as the co-evolution of visual and textual concepts.
6.1 Redefining Visual Storytelling with AESOP

In [42], crowd workers wrote natural language stories given a sequence of images from photo albums. Such a process leads to superficial and disjoint stories [180, 160] that focus on connecting text to image rather than on forming a coherent narrative. The limitation of such a process is evident when people are shown the story panels in random order. For over 35% of the stories, human observers are unable to find the “true” order of events, calling into question the value of such datasets for studying stories. Another limitation is that the dataset acquisition task is to generate text for a sequence of given images, with story writers having no control over the visual input. Consequently, a trained model is required only to produce a “feasible” text for a given image sequence.

Current visual storytelling research has dealt with tasks such as storytelling, generation [137, 42, 35] and illustration [3] or cloze tasks in [160] that primarily focus on cross-modal retrieval or generation. We discuss the limitations of such tasks and propose alternate tasks on AESOP that measure a system’s ability to comprehend and create stories from a true multimodal perspective requiring the perception and creation of both visual and textual modalities, that is absent in existing literature. The objective is for a system to be a creative assistant, by either autonomously or interactively assisting in creative processes like storytelling with visual, linguistic and narrative reasoning abilities.

Contributions

(1) AESOP,¹ a novel abstract visual storytelling dataset that captures the creative process associated with visual storytelling resulting in diverse, coherent and creative stories compared to existing datasets. Visual panels in our stories are composed by placing objects with various attributes in a scene, allowing for a broad range of expression and creativity. Each story is accompanied with annotations, including story theme and title, to promote future research on controlled or personalized storytelling.

¹Reference to Aesop, the Greek Fabulist and Storyteller.
Elaine was worried that John didn’t eat healthy. "John, would you like an apple instead? Maybe you should schedule a physical" she said. John told her to mind her own business.

That night John had a heartache on the living room. Elaine called an ambulance and John had to stay in the hospital for many days. He finally was well enough to come home.

John and Elaine bought bikes and ate lunch at the park sometimes. John discovered apples weren't too bad and Elaine was happy John was taking better care of them both.

Figure 6.1: Example story from our AESOP dataset with title and genres. The narrative is interesting, coherent and follows a clear causal arc with introduction and a moral at the end. The visual depiction of the story, including the changes in the expression of the characters, shows clear coherence and supports the narrative.
We propose novel story comprehension tasks on AESOP that demands multimodal, abstract, creative and causal reasoning ability from visual storytelling systems. Further, we propose a novel generalized story comprehension framework that models stories in our dataset as the co-evolution of visual and textual concepts.

(3) We quantitatively and qualitatively compare the proposed method and tasks with existing baselines and motivate our design choices through comprehensive ablation study. To the best of our knowledge, ours is the first work to study stories by aligning abstract visual and textual concepts and propose a comprehensive dataset, task and model to study important factors that govern visual storytelling, namely abstraction and creativity. We make the code and dataset publicly available to promote future research in this promising and challenging field of study.

6.2 AESOP Dataset

6.2.1 Guiding Principles

AESOP is built with the following three guiding principles.

Creativity Over Perception

Treating storytelling as merely a perceptive process limits creativity, inhibits diversity and result in stories that show sup-par temporal and causal coherence [3, 180]. In VIST, the ‘stories’ are written given semi-randomly chosen sequences of images. In AESOP, we ask crowd workers to create both the visual and textual parts of a story simultaneously from scratch, giving a lot more freedom for creative expression. We also limit the requirements, instructions and constraints to encourage creativity in the authors. As a result, the story on the whole conveys the semantic meaning as intended by the creator in contrast to the meaning as perceived by a viewer. We observe that the stories in our dataset are significantly more coherent with clear causal structure and more diverse.
Causal and Coherent Narratives

Stories are at minimum a causal sequence of events described in a coherent manner. For multimodal stories, such as ours, the need for coherence extends beyond just text. Since the stories in AESOP are created entirely from scratch instead of relying on prompts, they also exhibit themes with narrative arcs. To capture these, we also ask the story creators to provide each story with a title and genre (selected from a predefined list). Among other things, this can enable the training of models to produce genre- and title-conditioned stories.

Constrained World Knowledge

Comprehending stories using real-world images requires modelling the vast amount of implicit real-world knowledge represented in the images [14]. We seek to limit the complexity of the worlds our authors can create by simplifying the visual palette available to them. Inspired by [75, 76, 74, 73], we choose a clip–art based scene representation to depict the stories. As outlined in [75], usage of clip–art objects bypasses the step of object detection, localization and instance segmentation that would otherwise be required. Even with the visual simplifications, the diversity and creativity of stories and their accompanying illustrations are exceptional (refer supplementary materials). Moreover, since visual scenes in stories in our dataset are composed of clip–art objects, we can study visual story creation and comprehension that is not as easily possible in existing datasets like VIST [3].

6.2.2 Data Acquisition

Workers from Amazon Mechanical Turk authored our stories using a web interface that is an extension of the drag-and-drop tool used to generate the ‘abstract scenes’ portion of the VQA dataset [181], which is, in turn, an extension of the tool in [75]. We extend the number of clip art primitives from 149 to 158 and add two new backgrounds kitchen and beach, in addition to the park and living room backgrounds. Unchanged from [181], scenes in AESOP consist of 20 human characters with deformable limbs representing various ages,
Figure 6.2: All the clip–art objects present in the dataset along with the four backgrounds are shown for reference. The objects are all scaled to a uniform size for display (This causes blurring of small objects like bee). The actual sizes depend on what the object is and is different than what is shown here.
genders and races with 9 different possible expressions for each, and 30 animals and birds with various fixed poses for each animal. With our new object additions, it now includes 48 unique large objects related to outdoor and indoor scenes including sun, cloud, sofa, TV etc. and 60 unique small objects such as ball, cup, pizza etc. The large and small objects can also have sub-types depending on the type of object. In total, there are 158 unique objects that make up the visual parts of the story. Our final tool allows choosing and changing background, dragging objects onto the canvas, changing size, type, and depth of these objects and changing limb positions of each human figure. To ensure that the scene can be accurately reproduced from the story, we provided fixed names for each human figure which the workers were asked to use. (They were also free to use common nouns such as ‘a old man and his daughter’ instead.) We enforce some minimum constraints to dissuade low-effort submissions. First, the stories must contain at least one human in each scene so that the stories are human-centric. We also require a minimum number of changes between scenes so that not all the visual panels are identical. In addition to the visual story, the workers are also asked to provide a suitable free-form title and choose multiple themes from a list of predefined themes.

All the backgrounds and clip-art objects present in the dataset (excluding types) are shown according to their categories in Figure 6.2. Objects are scaled to the same size for viewing and might blur small objects like butterfly and bee. The actual sizes of the objects and their maximum and minimum depends on what the object is and can be changed within a fixed scale using the web interface while creating the story. The clip-art objects that were not present in [182] are highlighted in blue including the two new backgrounds. In total we have 159 unique objects and 291 total instances of all objects that includes subtypes. Types of objects may indicate different positions, colors, perspectives etc. For example, Fridge and Microwave can be open or closed making two types, while Window has three colors and three shapes giving nine types in total.
Figure 6.3: Top half of the data collection interface showing the instructions to workers followed by some examples of good stories. Important and necessary instructions are highlighted in red. Otherwise, constraints are limited and the workers are asked to be creative.
Figure 6.4: Remaining part of the data collection interface showing a preview of the story in its current form (followed by canvas to create visual panel and then space to provide story text. To the right of the canvas are all the clip-arts split into categories that can be dragged to the canvas.
Figure 6.5: Once the stories are written, workers are asked to select the genres for the story they wrote from a list of predefined genres. They are also asked to provide additional comments if available.

6.2.3 Data Acquisition Web Interface

The web interface used for data collection is shown in Figure 6.3 and Figure 6.4. Figure 6.3 shows the top half that has the instructions given to workers and some examples of ‘good’ stories to inspire them. The examples are chosen randomly from a set of stories each time a worker starts a HIT. The instructions were minimal with focus on coherence, character names, non–offensive stories and creativity. The choice to not condition the story creation process on any kind of additional input results in creative and diverse stories. The bottom half of the web interface is shown in Figure 6.4 and displays preview of the story being created along with the title. The current panel being edited is shown below the story preview. It contains an empty canvas to the left and all clip–art objects under four categories in the
right. Workers can change scene type to change the background, change orientation (flip) of each object and its z value by using the slide par below the canvas. An example of ‘sub types’ is shown in Figure 6.4 where the TV has four different types as can be seen below the canvas. The text corresponding to the current panel is written in the space provided at the bottom. Once the current visual and text parts of the story are completed, workers can continue on to the next panel. They have an option to copy the previous scene and start from that instead of from scratch for successive visual panels. The genre is asked when workers submit the story as shown in Figure 6.5.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Per story</th>
<th>Per Panel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg # unique objects</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Avg # unique humans</td>
<td>2.38</td>
<td>2.04</td>
</tr>
<tr>
<td>Avg time to create visual (in seconds)</td>
<td>950</td>
<td>–</td>
</tr>
<tr>
<td>Avg # words</td>
<td>84</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 6.1: Object and Word statistics of AESOP dataset.

6.2.4 Data Statistics

Stories in AESOP are made of 3 image-text panels with the visual parts generated by the drag-and-drop interface described above. We collected a total of 7,062 stories making up 21,186 abstract visual scenes and corresponding text created from scratch. Some word and object statistics are shown in Table 6.1.

Distributions of objects across stories in the dataset according to their categories is given in Figure 6.6 and Figure 6.7. Objects like sun, cloud, tree, bush are some of the most common and found in most of the park or beach scenes whereas Window is the most common object that concerns indoor scenes. Most common animal is dog while Bluejay is not used in any of the stories. Similarly, Basket, DishesMeal that has varieties of food on plates are the most common small objects. In humans Bobby and Lizzy of kid age group are the most common while Jack and Eddy that are toddlers are the least common.
Figure 6.6: Distribution of objects in visual panels across the entire dataset for Animals (Top), Large objects (Bottom).
Figure 6.7: Distribution of objects in visual panels across the entire dataset for Small objects (Top) and humans (Bottom).
6.2.5 Example Stories

The design principles explained in section 6.2 in the main paper has resulted in diverse and creative stories in AESOP. More examples from the dataset are given in Figure 6.8–Figure 6.11. For example, there are stories based on fantasy genre such as the third story in Figure 6.8 about a dystopian future or the fourth story in Figure 6.9 about a magic door. There are stories with moral such as second and fourth stories in Fig. Figure 6.8 and second story in Figure 6.9, or stories that emulate day–to–day activities (first story in Figure 6.8 or third in Figure 6.10).

There are also stories where objects are used in situations that do not usually define the characteristics of that object. For example, we have a significant number of stories based on COVID–19 where seashells are used as masks, starfish on a person’s shirt for indicating they are police, stapler as a gun, pond as portal, CD as wheels and so on. Some examples of such stories are shown in Figure 6.11. We expect that reasoning about such visuals require modelling the visual appearance such as regular pixel image features obtained from pre–trained CNNs. As part of future work, we would like to explore adding pixel information along with the abstract representation used in the models to overcome this limitation.

6.2.6 AESOP Vs. Other datasets

We comprehensibly analyze AESOP to study how it overcomes the limitations of existing vision–language and visual storytelling datasets.

Diversity

Verbs in text can be used to provide a notion of diversity in a dataset [180]. Compared to VIST [42], MSCOCO [12] and Flickr [183], the AESOP dataset shows more frequent use of verbs (Table 6.2). Furthermore, verbs in our dataset are also more diverse and longer-tailed with top-30 verbs providing a much smaller percentage coverage compared to these datasets. Following [180], we use the the American National Corpus ANC [184]
Alice bought a new bone for her dog. She wanted to bury the bone in her backyard so her dog would have to dig to find it.

Alice went up to her dog and told him that she had a bone for him. "Go get the bone!," Alice told him.

The dog was not interested in the bone. He just wanted to play with Alice.

On a cold winter’s day, Harry found a half frozen hawk. Having pity, he took it home to warm it up.

But when it warmed up, it attacked his family. Gratitude is not to be expected from the wicked.

The Earth was moving closer to the sun. Life was becoming extinguished. Alice fell to the ground, unable to take the heat anymore. She passed out.

Alice awoke in her bed, a doctor and a neighbor around her. The doctor said her fever had gone away and she’d be alright now. It was all a dream.

But she heard her friends talking. The Earth was moving away from the sun. If it didn’t stop, everyone would freeze.

Ryan was teaching his son Bobby how to ride a bike. Bobby was nervous as it was his first time with a real bike, not the ones with wheel attachments that he is used to.

Bobby fell off the bike. While Ryan is worried about him, he pretends nothing happened. "Try again, don’t let anything stop you," Ryan told Bobby.

Bobby tried again. This time he succeeded, he learned that mistakes are part of learning; you have to keep trying until you do it.

Figure 6.8: Example stories from AESOP dataset with title and genres. The changes in location, pose and expression of objects align with the events in the story. Also events in the story have clear causality and coherence with a diverse set of backgrounds, poses, scenes and story arcs.
"We're late, we need to have a quick dinner", said Mike. "I'm going to play for you and you pass it on to our daughter, she sets up dinner". He completed.

They started to pass the ingredients on when their daughter was dismayed to see her doll on the kitchen floor. "Look, my doll is here"

The couple's daughter dropped everything on the floor and stayed at home punished. She needed to be grounded to learn to pay attention to things.

Lizzy and Jack are Alice's Kids. Jared is her step son. Lizzy hates his step brother. Alice is very partial and gets most of the household chores done by Jared. On a sunny weekend Lizzy forces her mother to take her to the beach side and have fun.

Alice takes Lizzy to the beach and plays frisbee disc while Jared has to take care of her young kid Jack. Alice gets a sprain while jumping and the frisbee falls on the sea. Lizzy runs towards the sea to pick the frisbee up and gets in to a wave.

Jared leaves Jack on the towel, runs with the lifebuoy and saves Lizzy from being pulled in to the sea. Alice and Lizzy understands Jared's good heart and affection to his step sister, regrets for mistreating him and started to love him thereafter.

Colin was walking at the park when two young men, much bigger than him, stopped him and said "Hey boy, pass the cell phone, otherwise we will hit you". But Colin was without a cell phone, so the two went after him.

What they did not know is that Colin was raised alongside monks, masters of kung fu. And he started to defend himself against both, applying all the years of training.

Colin then left, leaving the two men injured on the floor.

Lizzy was walking across the beach, everything was fine, she went walking for hours when suddenly she saw a door. There was nothing supporting it. It was like the door was standing up because of pure magic.

Lizzy entered through the door and then appeared at the park. It was magic! she thought. She was very excited, she just had discovered a magical door.

When she tried to go back to the beach through the door, it just disappeared, like it was never there. Lizzy got trapped in this new city she didn't know.

Figure 6.9: Example stories from AESOP dataset with title and genres. The short and interesting titles and the genres indicating the entire emotional arc in the story are useful for emotional perception and controlled storytelling tasks. They could also be provided as additional input to the model to condition the generation of missing panel or even entire stories.
It was a warm summer day. My mom decided that she needed a day off of work because she was working very hard on a big deal she was hoping to get. She thought that we needed a day of fun and decided that a day at the park would be perfect for us.

After spending two hours at the park we decided to walk to a clear area and have a picnic. We ate fruit and had some snacks. The sun was hot but we sat under the tree for some shade. We talked about what a fun and enjoyable day we are having.

After a great day at the park, we went home and made ourselves a feast of hotdogs, cheese and popcorn. My puppy Gracie was so glad to see us because she doesn’t like being home alone. Mom and I are so happy that we got to spend the day together!

It was the first day of the school. Steve advised his children Bobby and Carol to be attentive in the class and to do everything, as directed by the teacher.

When Steve came back from office he was surprised to see Bobby sitting on his pet dog and writing something on his class note. "What are you doing?", Mike asked. Bobby replied, "Teacher told us to write an essay on our favourite animal, which I am doing"

Outside the house in the garden he saw his other child Carol. She was sitting on a turtle writing the essay. Turtle was her favorite animal. Steve shook his head and said to himself, "Now there is no doubt that both these children are really mine!"

It was a rainy day when Carol was told by her grandpa Jeff to help him to cook. Carol was very happy because she always wanted to cook with him.

After cooking lunch, Carol asked his grandfather if they could also cook a cake. He accepted and so they cooked a chocolate cake.

After that they laid the food on the table and ate with the whole family.

Mike was watching tv when suddenly an unusual promotion appeared. The advertisement was about buying your own clone.

Mike was curious about the commercial, so he did what every bored man would do in his position; a bad decision.

And so he did, he called for a clone, and there it was, another version of him. It was true; he thought that it was a joke; he was shocked. Next time he would think twice about buying something on tv.

Figure 6.10: More examples from AESOP dataset. we have stories that talk about clones, normal day of cooking, outing with mom and even a funny story with the title ‘Dumb and Dumber’.
I have finally made it to the portal. I need to close it before more aliens transfer through.

They are multiplying so fast. I will now repeat the words and use the magic wand to seal the portal.

Grumpy Trumpers, Grumpy Trumpers, seal the south forever and ever. Thank god it worked. May the south never rise again.

The South Shall Rise
(Genres: Funny, Drama, Fantasy)

Ryan was under quarantine due to Covid-19. His wife Emily and daughter Lizzy left to her parent’s house. Ryan was alone not knowing what to do these 14 days. He had an idea.

He planted many plants, transplanted many bushes and literally converted the front portion of his compound to a beautiful garden. He even created a fish pond too.

When his wife and daughter came back they were very happy and astonished to see the beautiful garden created by him. He understood that there is a silver lining in every cloud.

Silver Lining in the Cloud
(Genres: Happy, Moral, Drama)

Colin and his mule were driving a heavy wagon load of wood through the forest.

Colin was sick of hearing the screeching wheels. He noted that the mule was silent, so why couldn’t the wagon be quiet, too?

Colin gave his mule an apple, noting that those who cry the loudest are often the least hurt.

Screeching Wheels
(Genres: Moral)

Figure 6.11: More examples from AESOP dataset where objects are used creatively for purposes that does not usually define the object such as CD for wheels, seashell for masks and so on.
Table 6.2: Comparison with other datasets. Verb Frequency is the percentage of verbs over all words in the text. Top 30 verbs is the percentage of top 30 verbs over all verbs. Visible and Non-visible verbs indicate the frequency of select words per million words.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Verb Freq.</th>
<th>Top 30 Non-Visible Verbs</th>
<th>Visible Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Worry</td>
<td>Wonder</td>
</tr>
<tr>
<td>AESOP</td>
<td>0.198</td>
<td>0.589</td>
<td>556.0</td>
</tr>
<tr>
<td>VIST</td>
<td>0.017</td>
<td>0.669</td>
<td>9.8</td>
</tr>
<tr>
<td>MSCOCO</td>
<td>0.026</td>
<td>0.724</td>
<td>0.1</td>
</tr>
<tr>
<td>Flickr</td>
<td>0.012</td>
<td>0.723</td>
<td>0.1</td>
</tr>
<tr>
<td>ANC</td>
<td>0.184</td>
<td>0.563</td>
<td>143.6</td>
</tr>
</tbody>
</table>

for reference to what we can expect from a “natural” text. We can clearly see that AESOP most closely resembles the distribution and frequency of verbs in ANC. Furthermore, if we look at the characteristics of the verbs in existing datasets, most of them are visible verbs [180] that have visual grounding like *sit* and *talk*. Though this is understandable in the context of image captioning, it is undesirable in a storytelling VIST dataset. We believe this is due to the acquisition process being perceptive in nature. On the other hand, AESOP has more affective and non-visible verbs such as *worry* and *wonder* as there are no constraints on the creative flow in visual storytelling.

**Coherence and Causality**

To establish the extent of causality and coherence in AESOP compared to VIST, we perform a user evaluation where we asked humans to pick the correct story between the ground truth and a jumbled version of the story for 500 randomly chosen stories from both datasets. In the jumbled version of the story, two consecutive panels are swapped (excluding the first panel). Only 65.8% of stories from VIST were identified correctly while 95% of stories in our dataset were identified correctly showing that stories in our dataset have clear causality and coherence. An example story from each dataset is shown in Figure 6.12. It is hard to tell the correct order of panels in the VIST story, whereas in the AESOP example it is clear
Figure 6.12: An example story from VIST (top) and AESOP (bottom) with two consecutive panels highlighted in Blue. Swapping the highlighted panels in VIST gives a story that is indistinguishable from the original showing lack of causality and coherence. In our dataset, swapping these panels would lead to a meaningless story.

that pain from the dance is a result of the dance indicating clear causality.

6.3 Towards Comprehension and Creative Composition of visual stories with AESOP

We describe two well-defined tasks that take preliminary steps towards the grand goal of creating models that are truly capable of comprehending and creating stories. We posit that a fundamental requirement of such a model is the ability to continue and conclude a story started by a human. This setup, while being easy to train and evaluate, demands the models to maintain the consistency in the arrangement of objects and characters, and also be able to advance the story as suggested by the causal, motivational and narrative development in the prior story states. To this end, we define the following two tasks:
6.3.1 Assistant Illustrator

The Assistant Illustrator is required to generate the missing visual panel given the other two visual and all three textual panels. The aim of this task is to condition the visuals on existing panels while still measuring its ability to be visually reasonable and coherent as function of the input story. This can also be thought of as a generative variant of image-cloze task discussed in [160]. A human baseline example for assistant illustrator is shown in Figure 6.14. Even though many possible scenes can satisfy the story constraints, the objects and characters that are grounded in story often share consistent location, expressions and poses, unless explicitly mentioned in the text, making the original illustration a reliable ground truth for training purposes.

6.3.2 Assistant Writer

This is the text-equivalent of Assistant Illustrator where one of the textual panels is masked and the model completes the story by generating the missing text. This way, stories are grounded by some context to make evaluation more reasonable in contrast to Visual Storytelling [42]. A human baseline example for this task is shown in Figure 6.15. Note how the text is semantically similar to the original as a result of the conditioning on other visual and textual panels. This could also be thought of as a generative multimodal variant of the story cloze task [22, 160]. We believe this task ensures models rely explicitly on causality and cross-modal coherence compared to visual storytelling as the generations are not open-ended with no story specific context.

We note that even without additional annotations, AESOP can support various other tasks such as cross-modal generation instead of completion et cetera. As models make progress in the above tasks, we envision the creation of various new tasks using AESOP, fueling the development of models that can tackle more challenging storytelling tasks, making strides towards the creation of a truly intelligent and creative assistant.
6.4 AESOP The Model

Following the approach of [76, 74, 73], we treat visual panels as a sequence of objects and attributes. Our overall model is shown in Figure 6.13.

6.4.1 Abstract Visual Representation

We encode each visual token (an object) by encoding what the object is, where it is placed, and how it is placed to represent the state of that object. A visual panel is represented as \( V = [v_0, v_1, v_2, \ldots, v_{n_{max}}] \) where each \( v_i = (o_i, x_i, y_i, z_i, flip_i, pose_i, expr_i) \). We fix \( n_{max} \) to be a maximum of 15 in our experiments. Hereafter, we refer to \( n_{max} \) as just \( n \) for ease and each panel can have a varying number of objects less than or equal to \( n \). Here \( o_i \in [0, 290) \) is the object identifier, \( x_i \in [0, 700), y_i \in [0, 400) \) gives the location of the center of the object in the panel, \( z_i \in [0, 5) \) indicates size of the object, \( flip_i \in \{0, 1\} \) indicates whether the object is facing left or right, \( pose_i \in [0, 20) \) is the pose and \( expr_i \in [0, 10) \) indicates one of the nine possible expressions for human clip-arts. The first token \( v_0 \) indicates one of the four possible backgrounds added to the object vocabulary. Its attributes are all 0s. For human pose, we cluster the deformable rotation values (in radians) of the 9 independent parts such as torso, top and bottom arms, top and bottom legs for both left and right sides using \( K\)-means clustering [185] over the entire training set. We empirically fix 20 as the number of poses and ensure it covers most of the scenarios in the data. Though it would be better to predict the rotation values directly to model variety and creativity, pose estimation and generation are hard problems [186] and out of the scope of this paper. The order of objects is decided by the order in which they are placed on the scene by the renderer to create a scene [74, 73]. This ensures farthest objects like sun, cloud, boat are rendered first.
Mike tried to retrieve the ball but tripped and fell into the pond.

Mike was playing football with his best friend in the park one day. He unknowingly hit the ball behind his friend into the pond. Mike tried to retrieve the ball but tripped and fell into the pond.

Figure 6.13: AESOP model architecture containing a Text and Panel Encoders, followed by cross-modal attention and hierarchical decoders to generate a visual panel. (Zoom in for details)
followed by other objects. Each object is then encoded as

\[
\begin{align*}
\text{what}(v_i) &= LN(o_{emb}(o_i) + g(\text{word}(o_i))) \\
\text{where}(v_i) &= LN(f_{loc}([x_i; y_i; z_i; \text{flip}_i])) \\
\text{how}(v_i) &= LN(p_{emb}(\text{pose}_i) + e_{emb}(\text{expr}_i)) \\
f(v_i) &= \text{what}(v_i) + \text{where}(v_i) + \text{how}(v_i),
\end{align*}
\]

(6.1)

where \(o_{emb}, p_{emb}, e_{emb}\) are embedding layers similar to word embedding layers, \(LN\) is the layer normalization and \(f_{loc}\) is a linear layer. Values \(x_i, y_i, z_i, \text{ and flip}_i\) are normalized to be between 0 and 1 before embedding. We tried using embedding layers for location values as well similar to [73] but obtained better performance with this approach.

### 6.4.2 Story Encoder

Let \([V^1, V^2, V^3]\) be the sequence of visual panels that correspond to the sequence of text panels \([S^1, S^2, S^3]\). Then we represent the entire story using sequences of visual and textual tokens as \([f(v^1_1), ..., f(v^1_n), f(v^2_1), ..., f(v^2_n), f(v^3_1), ..., f(v^3_n)]\) and \([g(w^1_1), ..., g(w^1_n), g(w^2_1), ..., g(w^2_n), g(w^3_1), ..., g(w^3_n)]\) respectively where \(g(w^j_i)\) is the word embedding corresponding to the \(i\)th word in the \(j\)th text. For brevity, we lose the superscript that indicates the panel number and represent the entire story as a sequence of visual and textual tokens. To use the same model for all tasks, we simply replace the sequence of tokens responsible for the missing panel with a special \(\langle \text{MASK} \rangle\) token. Between the panels, we add a \(\langle \text{SEP} \rangle\) token and in the beginning and the end \(\langle \text{SOS} \rangle\) and \(\langle \text{EOS} \rangle\) tokens respectively.

The story encoder consists of a visual, text and a cross–modal encoder. The visual and textual encoders are separate Bidirectional GRUs [170], that encode modality specific coherence in the story. While the text encoder learns plausible story lines, the visual encoder learns plausible visual sequences. Next, we perform cross–modal attention between the encoded representations of the visual and textual tokens to provide cross–modal context. The story encoder consists of a visual, text and a cross–modal encoder. The visual and textual
encoders are separate Bidirectional GRUs [170], that encode modality specific coherence in the story. While the text encoder learns plausible story lines, the visual encoder learns plausible visual sequences as given in (Equation 6.2).

\[
\begin{align*}
    h_{v_i} &= Enc_{vis}(f(v_i), h_{v_{i-1}}, h_{v_{i+1}}) \\
    h_{w_i} &= Enc_{text}(g(w_i), h_{w_{i-1}}, h_{w_{i+1}}) \\
    h'_{v_i} &= \phi_{text}([h_{v_i}; g(word(o_i))], [h_{v_i}; g(w_i)]) \\
    h'_{w_i} &= \phi_{vis}([h_{w_i}; g(w_i)], [h_{v_i}; g(word(o_i))])
\end{align*}
\]

where \( Enc_{vis} \) and \( Enc_{text} \) are the BiGRUs, \( h'_{v_i} \) and \( h'_{w_i} \) are the final object and word representations. We initialize word embedding layer \( g(\cdot) \) with pre--trained glove embeddings [187]. The function \( \phi(\cdot) \) is the dot product attention layer similar to [104]. We concatenate the word embeddings associated with each object and word along with their respective learned representations before the attention layer. We do not provide separate names for each sub--type of the objects but use the general name to semantically ground the objects in text to avoid dependency on explicit object labels.

Figure 6.14: Examples of Assistant Illustrator result by Ground truth, Human Baseline, Proposed model and Unimodal are shown.
6.4.3 Panel Decoder

Visual Panel

We pose generation of the masked visual panel as \[74\] prediction of the following sequence \(V = [v_0, v_1, v_2, ..., v_n]\). We use two GRUs one to track the sequence of objects and another to track the state of the visual panel. The hidden state of both the GRUs are initialized with the final hidden states of the visual and text encoders. At each time step, the object decoder combines the state of objects predicted so far and attention over object and word representations from inputs, to predict the current object. Then the attribute decoder uses the predicted object along with current state of the scene to attend over objects in previous scenes and words in the text to predict attributes of the current object as a single \(33\)-dim vector, 4 for \(x_i, y_i, z_i\) and \(flip_i\), 20 for poses and 9 for expressions. The dimensions corresponding to \textit{where} attributes are clamped to be between 0 and 1 while \textit{softmax} function is applied for pose and expression classification.

The hidden state of both the GRUs are initialized with outputs of \(Enc_{vis}\) and \(Enc_{text}\) as \([h_{v_0}; h_{v_n}] + [h_{w_1}; h_{w_n}]\). At each time step, we first predict what is the next object based on the objects added to the scene so far as given in (Equation 6.4).

\[
\begin{align*}
    o_{i}^{state} &= Dec_{obj}(what(v_{i-1}), o_{i-1}^{state}) \\
    o_{i}^{vis} &= \phi_{obj}^{vis}(o_{i}^{state}, [h'_{v_0}, ..., h'_{v_n}]) \\
    o_{i}^{text} &= \phi_{obj}^{text}(o_{i}^{state}, [h'_{w_0}, ..., h'_{w_n}]) \\
    o_{i} &= MLP(o_{i}^{state}, o_{i}^{vis}, o_{i}^{text}) 
\end{align*}
\]  

(6.3)

where \(Dec_{obj}\) is a GRU that tracks the objects, \(g(o_{i-1})\) is the word embedding corresponding to the previous object, \(o_{i}^{state}\) is the current state of the object GRU. \(\phi_{obj}^{vis}\) and \(\phi_{obj}^{text}\) are linear attention layers similar to [170]. \(\phi_{obj}^{vis}\) attends to input visual panels to model visual coherence. Since story text is abstract, lot of the objects in the visual panel do not have explicit mentions in the text. Hence, to maintain coherence, it is imperative to have inde-
pendent attention over other visual panels. $\phi^\text{text}_i$ is required to attend to relevant text that is not visualized yet. Then, $o_i^{\text{vis}}$ represents suggestions based on other visual panels while $o_i^{\text{text}}$ represents suggestions based on text. Finally, the object is predicted by combining the current object GRU state and the visual and textual object suggestions. We treat the object prediction as a classification over the entire object vocabulary. Next, we decode the attributes of the predicted object as follows.

$$V_i^{\text{state}} = \text{Dec}_{\text{attr}}(\text{what}(v_{i-1}) + \text{where}(v_{i-1}) + \text{how}(v_{i-1}),$$

$$V_{i-1}^{\text{state}})$$

$$V_\text{state} = [V_i^{\text{state}}, \text{what}(v_i)]$$

$$\text{attr}^{\text{vis}}_i = \phi^\text{vis}_{\text{attr}}(V_i^{\text{state}}, h'_{v_j})$$

$$\text{attr}^{\text{text}}_i = \phi^\text{text}_{\text{attr}}(V_i^{\text{state}}, h'_{w_j})$$

$$\text{attr}_i = \text{MLP}(V_i^{\text{state}}, \text{attr}^{\text{vis}}_i, \text{attr}^{\text{text}}_i)$$

(6.4)

where $I_\text{state}$ is the current state of the visual panel with all the objects and their attributes so far. We combine the representation of the predicted object $o_i$ with the scene state as query for visual and textual attention. The visual and textual attention modules $\phi^\text{vis}_{\text{attr}}$ and $\phi^\text{text}_{\text{attr}}$ suggest possible set of attributes for the predicted object based on the current scene state and input panels. Finally, the attributes are predicted as a single 33–dim vector, 4 for $x_i$, $y_i$, $z_i$ and $\text{flip}_i$, 20 for poses and 9 for expressions. The dimensions corresponding to where attributes are clamped to be between 0 and 1 while softmax function is applied for pose and expression classification.

**Text Panel**

To generate missing text panel, we simply replicate the object decoder from the visual panel generator. Only modification is the vocabulary size for final classification of the word. The text panel decoder is trained using regular Maximum Likelihood objective.
The girls went to the beach for a ballet class. It was their first time taking a beach ballet class.

Harry has a chair on the beach so he sat on the chair and watched them train. The women played very good to add their self in the school. It was a great day for all. The weather is so cool and the day so suitable for exercising that the three of them end up smiling and practising their favorite sitting yoga poses.

They started practicing their poses. They followed the lead of the instructor. The students chuckled after the last move because they realized how out of shape they were! That made them very tired!

Figure 6.15: Examples of Assistant Writer result by Ground truth, Human Baseline, Proposed model and Unimodal are shown.

During inference, nucleus sampling [188] is used to generate the final text.

6.5 Experimental Setup

6.5.1 Baselines and Comparison

Since there are no directly applicable existing techniques that we can compare against, we compare against baselines and ablated versions of the proposed model.

Repeat:
Most visual scenes have slight changes in pose and expression while the majority of the background objects remain the same. Hence, we evaluate a baseline that simply copies the previous panel to the missing one for Assistant Illustrator. This model is not applicable to the Assistant writer mode as text changes considerably between panels.

Unimodal:
Visual unimodal model excludes the text encoder, cross-modal encoder and the text de-
coder attention modules. For text, we fine-tune a pretrained GPT-2 model [20] on in-filling task [189], to generate the masked text.

**One-to-One:**
To show the effect of modelling stories as a sequence of events, we also train a model that generates the masked visual/textual panel given the textual/visual panel independently without story context.

**Pixel Model:**
In this model, the abstract visual representation in the proposed model is replaced with a pretrained ResNet–18 [168] network and visual attention modules perform spatial attention similar to [74]. We fine-tune the ResNet–18 encoder along with the overall model.

**Human Baseline:**
We ask human workers to perform the same tasks for a human baseline.

### 6.5.2 Evaluation

Given the subjective and abstract nature of the storytelling task, it is unclear how to design automatic metrics that can faithfully quantify a system’s ability to create or comprehend a story. However, to support fast prototyping and give a rough sense of correctness of predictions, we use the following metrics for the tasks.

**Illustrator Metrics**

Following the works of [76, 74, 73] we use accuracy of prediction for $o_i, z_i, flip_i, pose_i$ and $expr_i$ and background. For location we use the Absolute Similarity from [73]. Scene Similarity metric proposed in [76] is used for overall score, but treat pose and expression as ‘full’ targets instead of weighed by 0.5. We emphasize these factors because variations in pose and expression convey significant subjective story content (in contrast to descriptive scenes as used in [76]).
**Writer Metrics**

For text generation, we use existing metrics BLEU-k, METEOR, CIDEr and ROUGE-L [190].

**User Study**

Though the proposed completion tasks are more constrained than generic open-ended storytelling tasks, automatic evaluation based on absolute metrics is nevertheless unreliable due to ambiguity (consider, e.g. the human baseline in Figure 6.14). Hence, we have performed extensive user studies to compare the results of different baselines to more fairly assess the models. Specifically, we sample 500 random stories from the test set and ask humans to do the same task. An independent user group performs pairwise comparisons of each of the baselines, including the human baseline. Comparison is done along each of three dimensions, defined as follows:

1) **Coherent:** Is the generated content consistent with the preceding content? This mainly measures consistency within the same modality.

2) **Relevant:** Is the generated content relevant to the corresponding content from alternate modality in the same panel? This dimension measures the consistency between corresponding text and visual panels.

3) **Meaningful:** Is the generated content sensible? E.g., A meaningful representation of a living room will depict a sensible living room scene but may or may not show a good coherence with prior panels. This dimension measures the meaningfulness of the generated panel while ignoring the corresponding modality.

**6.5.3 Training Details**

Out of 7062 stories we use 5562 for training, 500 for validation and test on the remaining 1000 stories. All human evaluation experiments and human baseline models are run on a subset of the test set containing 500 stories. The hidden dimensions of the encoder and
decoder are 512 and the visual and text tokens have an output dimension of 1024 (including the word embeddings). Maximum number of words considered per story is 150, 50 per text panel, while the maximum number of objects is set at 45, 15 per panel. The maximum excludes special tokens such as \langle \text{MASK} \rangle, \langle \text{SEP} \rangle, \langle \text{SOS} \rangle and \langle \text{EOS} \rangle separately for visual and text. There is also \langle \text{SOG} \rangle and \langle \text{EOG} \rangle indicating start and end of generation for the decoders.

We use the Adam optimizer [174], initialized with a learning rate of 3.5e-4 and a decay rate of 0.8 whenever scene similarity metric plateaus with patience of 8 epochs. We train all the models for 80 epochs on the training set and the epoch with the highest Scene Similarity metric on validation data is chosen as the best epoch for evaluation on the test set.

During inference, metrics are calculated over the entire test set. We choose learning rate of 3.5e-4 was empirically chosen from \{1e-4, 1.5e-4, 2e-4, ..., 5e-2\}, hidden dimension of 512 for the networks from \{128, 256, 512, 768, 1024\} and batch size from \{8, 16, 24, 32, 48, 64\}. The word vocabulary size of the text encoder decoders is 11 158 while the object vocabulary is 291 + 10 for special tokens and backgrounds.

6.6 Results

6.6.1 Assistant Illustrator

Last Panel

We see from Table 6.3 that the simple ‘Repeat’ baseline gives higher scores for all metrics compared to the full model or even human baseline when using automatic metrics for scene similarity. This is mainly because for over 80% of the stories in the dataset, the background is unchanged. Moreover, many scene objects do not change their position or attributes throughout the story. We perform pairwise comparison of 4 models using human judges to further understand the reliability of the quantitative metrics and truly evaluate the performance of these models. The results are shown in Table 6.4. In contrast to the the observation in Table 6.3, we can see in the user study that human baseline clearly outper-
<table>
<thead>
<tr>
<th>Model</th>
<th>BG (\uparrow)</th>
<th>O-IOU (\uparrow)</th>
<th>Loc (\uparrow)</th>
<th>Dep (\uparrow)</th>
<th>Flip (\uparrow)</th>
<th>Pose (\uparrow)</th>
<th>Expr (\uparrow)</th>
<th>Scene (\uparrow)</th>
<th>B–1 (\uparrow)</th>
<th>B–4 (\uparrow)</th>
<th>M (\uparrow)</th>
<th>R–L (\uparrow)</th>
<th>C (\uparrow)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Illustrator</td>
<td>90.1</td>
<td>66.5</td>
<td>0.73</td>
<td>92.2</td>
<td>89.2</td>
<td>30.4</td>
<td>41.3</td>
<td>4.1</td>
<td>26.28</td>
<td>1.96</td>
<td>9.02</td>
<td>22.1</td>
<td>17.9</td>
</tr>
<tr>
<td>Unimodal Illustrator</td>
<td>89.8</td>
<td>68.5</td>
<td>0.71</td>
<td>92.2</td>
<td>89.5</td>
<td>33.3</td>
<td>37.9</td>
<td>4.2</td>
<td>10.06</td>
<td>0.41</td>
<td>6.7</td>
<td>10.4</td>
<td>5.7</td>
</tr>
<tr>
<td>One-to-one Pixel</td>
<td>68.3</td>
<td>18.0</td>
<td>0.42</td>
<td>46.4</td>
<td>26.1</td>
<td>5.42</td>
<td>7.32</td>
<td>1.2</td>
<td>23.04</td>
<td>1.72</td>
<td>7.2</td>
<td>10.8</td>
<td>7.2</td>
</tr>
<tr>
<td>Human</td>
<td>52.3</td>
<td>15.7</td>
<td>0.25</td>
<td>21.5</td>
<td>10.6</td>
<td>3.54</td>
<td>5.12</td>
<td>1.0</td>
<td>8.62</td>
<td>0.73</td>
<td>5.4</td>
<td>7.98</td>
<td>4.32</td>
</tr>
<tr>
<td>Repeat</td>
<td>95</td>
<td>72.6</td>
<td>0.86</td>
<td>73.7</td>
<td>70.6</td>
<td>23.7</td>
<td>38.2</td>
<td>4.0</td>
<td>21.08</td>
<td>1.84</td>
<td>11.1</td>
<td>15.1</td>
<td>18.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>BG (\uparrow)</th>
<th>O-IOU (\uparrow)</th>
<th>Loc (\uparrow)</th>
<th>Dep (\uparrow)</th>
<th>Flip (\uparrow)</th>
<th>Pose (\uparrow)</th>
<th>Expr (\uparrow)</th>
<th>Scene (\uparrow)</th>
<th>B–1 (\uparrow)</th>
<th>B–4 (\uparrow)</th>
<th>M (\uparrow)</th>
<th>R–L (\uparrow)</th>
<th>C (\uparrow)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Illustrator</td>
<td>92.0</td>
<td>68.5</td>
<td>0.77</td>
<td>94.2</td>
<td>90.8</td>
<td>32.6</td>
<td>43.2</td>
<td>4.3</td>
<td>25.6</td>
<td>2.5</td>
<td>8.8</td>
<td>22.5</td>
<td>17.6</td>
</tr>
<tr>
<td>Unimodal Illustrator</td>
<td>91.5</td>
<td>71.7</td>
<td>0.77</td>
<td>94.1</td>
<td>90.9</td>
<td>31.9</td>
<td>44.0</td>
<td>4.6</td>
<td>8.2</td>
<td>0.36</td>
<td>6.2</td>
<td>11.1</td>
<td>5.1</td>
</tr>
<tr>
<td>One-to-one Pixel</td>
<td>69.0</td>
<td>18.5</td>
<td>0.40</td>
<td>42.5</td>
<td>32.4</td>
<td>8.1</td>
<td>7.72</td>
<td>1.4</td>
<td>22.6</td>
<td>0.95</td>
<td>7.0</td>
<td>12.7</td>
<td>8.1</td>
</tr>
<tr>
<td>Human</td>
<td>55.4</td>
<td>15.9</td>
<td>0.28</td>
<td>21.4</td>
<td>12.1</td>
<td>4.8</td>
<td>6.12</td>
<td>1.3</td>
<td>11.2</td>
<td>0.85</td>
<td>7.9</td>
<td>18.2</td>
<td>16.1</td>
</tr>
<tr>
<td>Repeat</td>
<td>95.8</td>
<td>79.5</td>
<td>0.90</td>
<td>73.5</td>
<td>71.1</td>
<td>25.1</td>
<td>31.3</td>
<td>4.4</td>
<td>28.4</td>
<td>4.9</td>
<td>12.5</td>
<td>28.3</td>
<td>20.1</td>
</tr>
</tbody>
</table>

Table 6.3: Results of all models on Assistant Illustrator and Assistant Writer modes when the TOP: last panel is masked and BOTTOM: middle panel is masked. For Assistant Illustrator, we provide accuracy over entire test set for prediction of BG (background) Dep (z value), Flip, Pose and Expr (Expression). Loc is the location similarity while O-IOU is the intersection over union between predicted and ground truth set of objects. Metrics for object attributes are calculated only if the predicted object is present in ground truth. Scene is the scene similarity metric. For Assistant Writer mode, B–1 indicates BLEU–1, B–4 is BLEU–4, M is METEOR, R–L is ROUGE–L and C is CIDEr.
forms the ‘Repeat’ baseline by a large margin. This underscores the need for more reliable automatic metrics for this complex task. Additionally, according to user study, we can see that even though our proposed model is better than simplest baselines, it is far behind the human level performance.

**Middle Panel**

We observe that the values for the metrics for middle visual panel generation are higher for all the models when compared to the scores for last panel generation. This is because, there is much less surprise in the middle visual panels compared to the last ones. More stories have different backgrounds and new objects in the last panel while the middle panel contains minimal changes. If stories introduce new characters in the middle panel those mostly do not leave the story in the last panel, requiring models to learn which of the available two panels (first or last) to copy from to beat the metrics. Note that pose, expression and other changes still do exist but the metrics give an overall notion of similarity to ground truth and since most objects do not change significantly, learning to replicate the other panels can lead to higher scores as can be observed from the scores of ‘Repeat’ baselines. Repeat baseline again outperforms on most of the metrics. Human baseline begins to close the gap on these quantitative metrics because there are less variations in the human created scenes as well for the middle panel. Note that we do collect human baselines for middle panel generation as well but do not perform user study to evaluate the performance of the models.

**Qualitative Examples**

Figure 6.16 and Figure 6.17 show more examples of last visual panel generation by the proposed model compared with human, ground truth and unimodal baselines where the proposed model performed relatively better than baselines as rated in the user study. For example, in Figure 6.17 (left) the proposed and human baselines got same relevance and
<table>
<thead>
<tr>
<th>Experiment</th>
<th>Meaningful</th>
<th>Relevant</th>
<th>Coherent</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>77.2</td>
<td>84.5</td>
<td>81.8</td>
<td>87.1</td>
</tr>
<tr>
<td>Proposed</td>
<td>6.6</td>
<td>7.1</td>
<td>7.0</td>
<td>7.6</td>
</tr>
<tr>
<td>No preference</td>
<td>16.2</td>
<td>8.4</td>
<td>11.2</td>
<td>5.3</td>
</tr>
<tr>
<td>Proposed</td>
<td>29.4</td>
<td>32</td>
<td>30.8</td>
<td>35.8</td>
</tr>
<tr>
<td>Unimodal</td>
<td>26.2</td>
<td>23.6</td>
<td>24</td>
<td>27.8</td>
</tr>
<tr>
<td>No preference</td>
<td>44.4</td>
<td>44.4</td>
<td>45.2</td>
<td>36.4</td>
</tr>
<tr>
<td>Human</td>
<td>68.8</td>
<td>85.2</td>
<td>80</td>
<td>86.5</td>
</tr>
<tr>
<td>Repeat</td>
<td>7.2</td>
<td>4.6</td>
<td>6.6</td>
<td>5.3</td>
</tr>
<tr>
<td>No preference</td>
<td>24</td>
<td>10.1</td>
<td>13.4</td>
<td>8.2</td>
</tr>
</tbody>
</table>

Table 6.4: Results of user study comparing models pairwise along three dimensions for Assistant Illustrator. Values are given in % and _overall_ indicates the overall preference between the two shown models.

meaningfulness score but human baseline won on coherence giving the overall score to the human baseline. In the proposed vs unimodal comparison experiment, the proposed model won on all fronts for this story. We can see that the model is able to learn cross–modality relevance and visual coherence. In all the examples the first two panels are given at the top and the last panel for all the baselines are given below them with colored borders indicating the model type. Each figure also gives a brief analysis on the generations.

In Figure 6.18 we show some failure cases of the proposed model. In most of the failure cases, the model is unable to figure out the changes and ends up replicating the previous panel or tries to change but is unable to completely create a new unseen scene. Based on these results, we highlighted next steps for better models in Sec. 9 in the main paper.
Lizzy was very lonely. Her favorite doll was all she had. She wished it were real.

All of a sudden, the doll started growing. It came to life!

"Lizzy, I am so glad to finally meet you", the doll said. "You've taken care of me so well".

Ryan and Alice were both drinking and having fun in Alice’s house. Alice warned him that he is drunk, as he was supposed to return home on a motorcycle. "My genetics are extremely resistant to alcohol" he said.

Ryan then (finally) decided to go home. On the way he lost control of the bike and hit a tree.

The next day, after returning from the hospital, Alice went to visit Ryan. He was lying on the couch with several broken bones. "You are completely irresponsible, Ryan. You could have hurt someone! Our relationship is over! Good luck!" she said.

Figure 6.16: Examples of Assistant Illustrator for last panel generation result by Ground truth, Human Baseline, Proposed model and Unimodal where the proposed model successfully generated relevant and coherent visual panels.

**Analysis:**

**Left:** The generated visual panel removes the doll, retains the woman and changes her expression correctly. Note that the human baseline and ground truth do not show clustered poses and hence looks more realistic while the generated visual panel has predicted the closest pose from the 20 possible poses (during training ground truth and input poses are clustered but general poses are shown here to retain the realism in data). Unimodal as expected does not know to remove the doll or change expressions indicating that the proposed model takes text into account.

**Right:** The story starts at Alice’s house but ends at Ryan’s place. The ground truth story does not have a change of scene in the third panel. The human baseline however, correctly captures the change of scene. In the proposed model’s generation, the scene change from park to a house, Alice’s presence and her angry expression are all captured perfectly when compared with ground truth. However, the model misses that Ryan is lying on the couch.
The boy took his dog to the park. The dog went running ahead. The boy didn't know why, but the dog ran straight to a sand pile near the trees. What could the dog be looking for?

The dog stopped and sat at the pile and began to bark. The boy knelt down for a closer look. He could see something sticking out of the sand. He found a shovel and started digging.

There was a chipmunk buried in the sand! The boy was able to free him by digging him out of the sand. The chipmunk ran away very fast so he wouldn't get buried again! The boy was happy he saved the chipmunk.

Lizzy and Carol spent some time in the park yesterday. The two of them had decided to have a monkey bar competition. After just two minutes, Carol got tired and fell on the ground. Lizzy jumped off the bars and came to her friend's aid. She was concerned about her fall. Luckily she was not hurt. She was happy that her friend was safe.

Figure 6.17: Examples of Assistant Illustrator for last panel generation result by **Ground truth**, **Human Baseline**, **Proposed model** and **Unimodal** where the proposed model successfully generated relevant and coherent visual panels.

**Analysis:**

**Left:** The generated visual panel is close to ground truth as well as human baselines. It retains the sand, but relieves the chipmunk and has similar pose and orientation to human baseline for Ryan and the dog. Except for the turtle that seems to be on top of Ryan, the overall scene is relevant to the text, coherent to previous visual panels and depicts a meaningful scene.

**Right:** The proposed model captures the change in expressions and relative locations correctly similar to ground truth or human baseline making for a reasonable illustration of the story. Moreover the model has learned to not vary the position of objects that are still such as the monkey–bars, bush, and tree.
Jared was hungry, so he went to the kitchen to get something to eat. His dog, Tiny, was hungry, too. Tiny tried to be a good little dog, hoping Jared would feed him.

Jared opened the fridge and noticed a can of pop on the door and a hamburger patty on the bottom shelf. Tiny noticed the hamburger patty, too. “That hamburger patty might be good with that Coke.” thought Jared.

When Jared bent down to grab the Coke, Tiny jumped up and grabbed the hamburger patty and ran to the other room with it. “Oh, no!” Jared exclaimed. “I wanted that hamburger!” But Tiny beat him to it. First come, first served!

Jane brought her new camera to the park to take nature photos. Jane snapped a picture of beautiful blue butterflies.

Then Jane crouched down to get a close-up picture of some interesting looking mushrooms.

Finally, Jane backed up to get a photo of the stately tree. But when she did, Jane tripped over the edge of the pond and fell backward, dropping her camera in the pond.

Figure 6.18: Examples of Assistant Illustrator for last panel generation result by Ground truth, Human Baseline, Proposed model and Unimodal where the proposed model failed to generate relevant and coherent visual panels.

Analysis:

Left: The ground truth illustration for the last panel visualizes the Jared and dog before the dog goes to the other room while the human baseline visualizes them in another room. The proposed model also takes them to another room but lack of any details on what the room is in the story makes it difficult for the model to place them in reasonable locations. However, the model still got all the relevant objects such as couch and fireplace for the living room and dog, Jared, hamburger and the soda for the story.

Right: The model illustrates falling down but did not capture where exactly Jane falls down which should have been near the pool. Predicting each of the attributes is a separate task in itself (e.g. spatial reasoning in abstract scenes formulated as a separate task in [73]), making the overall task complex.

6.6.2 Assistant Writer

Last Panel

In the assistant writer mode, we can see how the full model achieves better score than baselines including human baseline for BLEU and ROUGE-L scores. The proposed model with visual information, has explicit object and attribute embeddings that ensures no characters are missed in the text thereby getting higher scores for these metrics. However learning to generate coherent narratives while also being relevant to visual information is hard for the
model causing its METEOR and CIDEr scores to fall. In user study for the Assistant writer mode, we observe that both human and GPT–2 versions outperform the proposed model significantly. We believe this to be because of the difficulty in learning language modelling by our model from scratch on the relatively small dataset. It generated grammatically incorrect text making it less preferable. A natural improvement to the current model would be to initialize the text encoder–decoder parts with pretrained language models to ease the burden in learning language coherence from scratch.

**Middle Panel**

The proposed model has similar BLEU and ROUGE-L scores indicating it is able to capture the contents present in the corresponding visual panel to some extent while it struggles to form grammatically correct and coherent sentences. We observe these through examples as well (Figure 6.19), and we anticipate similar results as last panel generation if we run human studies. GPT2 on the other hand scores slightly higher on the middle panel completion across all metrics. Human baselines outperform across all metrics for middle panel text generation.

We also tried a single model trained to generate any missing panel but given the bias towards minimum changes in the middle panel and the last panel for most stories, the model ends up learning to replicate the ‘Repeat’ baseline. This further motivates the need for a model that encodes only change of objects and attributes in the visual side. This would also ensure better alignment with text as the text has explicit mentions of only what changes throughout a story.

**Qualitative Examples**

Additional examples for the Assistant Writer task for the last panel are given in Figure 6.19 and Figure 6.20.
Ryan took Bobby to the beach on Saturday and the two of them noticed that the beach had a new waterslide. Ryan had decided to go up the slide on Saturday because of a new wave. Ryan and Bobby had a good time at the beach and they were so happy and proud of listening to him. Bobby agreed and ran to the slide. He started playing on it and felt thrilled. He asked his dad to accompany him on the slide and was glad that his dad was with him. Bobby them went down the slide carefully and had a blast. He could not wait to go back on the slide again.

Jeff and Harry were playing beachball in their house. They were having a lot of fun. The dog ran into the waters and hit Harry with the ball. Harry felt terrible. Harry felt terrible that his dog was going to hurt him. He asked for help. The dog hit Harry on the face of his head which caused him to fall on the ground. The dog then jumped on Harry and had knocked him over onto the floor leaving him injured.

Figure 6.19: Examples of Assistant Writer for last last panel generation result by Ground truth, Human Baseline, Proposed model and Unimodal where the proposed model successfully generated reasonable text to complete the story.

Analysis:
Left: The proposed model scored equally for relevance and meaningfulness against human baseline while the human baseline won against coherence. The proposed model’s generation won in all metrics against the unimodal GPT2 model. The generated text is a reasonable ending for the current story.

Right: In this example the generated text by the proposed model captures the content with high relevance and coherence to the rest of the story achieving higher scores in the human evaluation.
Mother and daughter decided to spend the weekend at the beach and enjoy the sunny days. They created sandcastles and threw the beach ball around, and had fun making sandcastles. The sun was so close, they built a sandcastle and they enjoyed it a sandcastle. They built a great sand castle on a fun day at the beach. They saw a pile of sand, and decided to build a sandcastle. They were really good at building a sandcastle, it was very well done!

It was winter, Emma was sitting in front of fireplace. While she was making her warm she heard a voice of a dog at her door step. Emma had started to approach the door to see if the dog was living with her. Emma gave the puppy a big plate. She told the dog it was just some to her dog and she was happy with the dog. She brought it inside, where it soon warmed up and to her. She allowed the dog inside her home and gave shelter and food. Emma shared her fireplace with the little dog and both had fun together. Emma gave the puppy a big plate. She told the dog it alone her decided it was just some to her dog and she was happy with the dog. Emma had started to approach the door to see if the dog was living with her.

Figure 6.20: Examples of Assistant Writer result for last panel generation by Ground truth, Human Baseline, Proposed model and Unimodal where the proposed model generated irrelevant or incoherent text to end the story.

Analysis:

Left: The proposed model’s generation shows how it loses on coherence while trying to be relevant to the corresponding visual panel. GPT–2 generates much more coherent text and keeping with the context of the other text panels makes it preferable.

Right: This is another example of incoherent text generated by the proposed model. while the objects are perceived as we would like, the text is not comprehensible. Initializing the text parts of the model with pre–trained language models as pointed in future work would help overcome this limitation.
6.6.3 Model Limitations and Future Work

Though the proposed model is able to capture cross-modal relevance and visual coherence better than baselines, it is far from achieving human level performance. Even with an abstract and constrained visual world, the diversity and creativity in the stories make this a complex task. This is because human creators still act upon years of accumulated world knowledge to create each story, which is difficult to capture using generic models based on existing literature. The current model learns to copy from previous panels or create new scenes if required by text but struggles to populate new scenes (more examples in supplementary materials). A natural extension to our model is to add pretrained language or multimodal models to initialize the network for better language–vision alignment and to ease the burden in learning language coherence. Further, given the minimal changes between visual panels in the stories, it might be reasonable to model visual panel completion as predicting scene changes rather than absolute scenes. Additionally, adding a variational generative component that is conditioned on the state of the story would provide creative abilities to the model. We also plan to add title and genre information to the encoders to condition the story state on user-defined context.

6.6.4 Inadequacies of Automatic Metrics

AESOP has emphasized the inability of automatic metrics to capture true notion of correctness for stories by contrasting the user evaluation results with those in Table 6.3. We plan to provide a platform to perform human evaluation using the defined dimensions in a standardized manner similar to common challenges such as VQA [181] in language–vision literature.

6.6.5 Complexity of AESOP

Compared to closely related works such as [74, 73] for abstract scene generation, AESOP is highly complex. Tan et. al. in [74] consider scene generation on a dataset with descriptive
and grounded text and considerably fewer (58 vs 158 in AESOP) objects and scenes. Similarly Radevski et. al.[74], only require spatial location prediction of objects for the same dataset. In comparison, AESOP not only requires grounding deformable limbs, more objects, expressions and backgrounds but also require models to do so using non-descriptive, inexact text that do not directly refer to objects in the scene. (Instead of text in [74]: ‘Mike is holding a hotdog. Jenny is walking towards Mike’, AESOP has: ‘Mike is having a picnic with his friends’). On the text-side, AESOP shows similar spike in complexity compared to closely related story-text generation tasks [42, 22], where the requirements are either ill–posed [42] or framed as an easier retrieval setup [22].

### 6.7 Limitations and Improvements

With the introduction of the AESOP dataset, we have established a new frontier in abstract visual storytelling. The AESOP dataset together with the tasks and initial baselines explored in this paper have paved a way towards the development of models capable of not only comprehending and creating visual stories but also working alongside humans to create powerful visual narratives.

The rich annotations that we have collected in AESOP allows for creation of many other tasks beyond the two described in the main paper. These include panel generation from story-text (Illustrator–mode), VIST-style story generation using panels (Writer–mode), controllable story generation using different title/theme prompts etc. We also envision collection of auxiliary annotations that can enable tasks such as collaborative story-writing, story question–answering (Who is the main character in the story? How is Emily likely to feel after this?) and others. We hope such developments will make strides towards the creation of a truly intelligent and creative assistant for writers and illustrators.
CHAPTER 7
IMAGE–TEXT COHERENCE

What makes an image relevant to a text is complex: it depends both on the meaning of
the text and the user’s motivation for linking the text to visual content. Current image-text
joint understanding techniques, as evaluated on standard datasets like MSCOCO [12] and
Flickr30K [183], work only on factual text that describe the contents of an image. However,
human multimodal discourse also contains affective and subjective information [191, 192,
193] and can combine complementary text and imagery to tell a unified story [194]. This
chapter describes models of these broader associations between text and imagery.

[195] argue that when people communicate with visual imagery and text, they intend
their audience to recover a specific structure for the discourse and specific inferences in
context. They propose that these structures and inferences can be modeled using represen-
tations and algorithms informed by approaches to natural language (NL) discourse, par-
ticularly coherence relations. Coherence relations characterize the inferential links (such
as temporal, causal, and logical) that connect the content of text and imagery. [13] show
that coherence-aware description generation models outperform the existing coherence-
agnostic state-of-the-art models, since they are sensitive to the communicative intent of
image-text presentations.

7.1 Exploiting Cross–Modal Coherence for Text–to–Image Retrieval

We study the role of coherence relations in text-to-image retrieval. We hypothesize that
bringing in coherence relations [13] into the retrieval process, in contrast to personalities
defined in [138], should better improve the performance of text-to-image retrieval in a
more generalizable way. We build on [196] and [197] and introduce a new framework that
integrates coherence relations in text-to-image retrieval task by extracting features for each
Caption: The start of the race.

![Image](image_url)

Figure 7.1: Example retrieved image by the proposed Cross-Modal Coherence Model (right) vs Cross-Modal Coherence Agnostic Model (left) for input caption (top).

modality separately then building a lower-dimensional common representation space. Our proposed framework introduces a Coherence Aware Module that learns to predict coherence relations during training, and the same module is also applied during testing to further improve the retrieval performance. This module helps the retrieval model focus on the intent of the users and the potential effects of image–text combination.

Text-to-image retrieval models have been used in several multimodal NLP tasks and applications. [77] extract syntactic relations from captions for indexing and retrieving photographs of crime scenes. [78] use image retrieval as a testbed for learning spatial relationships between image regions using Visual Dependency Representations. Several works have shown that including images in information retrieval tasks such as document retrieval can improve the performance of the models [79, 80]. Most recent visual dialogue systems include image retrieval models to present images with text in response to user’s needs and to better fulfill the dialogue goals [81]. None of the previous works in this line have studied a discourse-aware approach for text-to-image retrieval which would best suit the context of the dialogue, inferences between text and imagery in multimodal documents, and the role
of coherence in learning better models of image-text alignments.

The majority of the previously proposed techniques propose a two-stage framework (e.g. [198]). In the first step, features for each modality are extracted separately, and then a lower-dimensional common representation space is built using canonical correlation analysis [199] or cosine similarity [196, 197, 200, 201]. These techniques assume a single overall relation over all image-text pairs. Although some works [202] leverage meta-tags to regularize features from different domains, they still do not explicitly model different coherence relations that characterize image-text pairs.

[195, 13] characterized coherence relations in text and imagery. They evaluated the effectiveness of coherence relations on a controlled caption generation task where an image and a coherence relation are given as input while the model generates a caption that adheres to both the image and the relation. We do not train a controllable model as we hypothesize that not all relations equally characterize the image and text in a pair. Though the relations are defined for joint image-text discourse, some coherence relations like “Subjective” in the Clue dataset characterize how the caption relates to the image and not the other way around. Hence, conditioning image retrieval on the relation is not reasonable. The proposed method evaluates the effectiveness of coherence relations by comparing \textit{CMCA} with \textit{CMCM}. Note that the proposed \textit{Cross-Modal Coherence Model} is not the same as in [13]. Instead, our model learns to predict the coherence relation during training. To the best of our knowledge, this is the first work that comprehensively analyzes the effect of coherence relations for image retrieval and trains text-to-image retrieval models with cross-modal coherence.

The examples of Figure 7.1 illustrate our approach. They contrast the output of our baseline \textit{Cross-Modal Coherence Agnostic Model (CMCA)} taken from [200] and that of the proposed \textit{Cross-Modal Coherence Model (CMCM)} trained on image-text pairs with \textit{Story} coherence relations. We observe that the proposed \textit{CMCM} provides more importance to the words \textit{the start} compared to \textit{CMCA} that concentrates on visually grounded words like \textit{race}. Thus, \textit{Coherence Aware Module} provides more interpretable and robust results, by virtue
<table>
<thead>
<tr>
<th>Relation</th>
<th>Question</th>
<th>Description</th>
<th>Positive rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expansion</td>
<td>Q2</td>
<td>The image gives visual information about the step described in the text.</td>
<td>0.821</td>
</tr>
<tr>
<td>ImageNeeded</td>
<td>Q3</td>
<td>You need to see the image in order to be able to carry out the step properly.</td>
<td>0.115</td>
</tr>
<tr>
<td>Elaboration</td>
<td>Q4</td>
<td>The text provides specific quantities (amounts, measurements, etc.) that you would not know just by looking at the picture.</td>
<td>0.329</td>
</tr>
<tr>
<td>Elaboration_{tool}</td>
<td>Q5</td>
<td>The image shows a tool used in the step but not mentioned in the text.</td>
<td>0.193</td>
</tr>
<tr>
<td>Temporal_{&lt;t}</td>
<td>Q6</td>
<td>The image shows how to prepare before carrying out the step.</td>
<td>0.158</td>
</tr>
<tr>
<td>Temporal_{&gt;t}</td>
<td>Q7</td>
<td>The image shows the results of the action that is described in the text.</td>
<td>0.588</td>
</tr>
<tr>
<td>Temporal_{=t}</td>
<td>Q8</td>
<td>The image depicts an action in progress that is described in the text.</td>
<td>0.313</td>
</tr>
</tbody>
</table>

Table 7.1: Coherence relations, their distribution and entropy in CITE++ dataset. We use the question identifier and the relation name interchangeably. *Positive rate* is the percentage of samples that are labeled as ‘Yes’ for that question

of explicitly modelling image-text coherence.

We propose CMCM with an auxiliary *Coherence Aware Module* that helps the model learn what type of coherence relation characterizes an image-text pair during training for text-to-image retrieval task. During inference, the model leverages this knowledge to retrieve relevant images. In section 7.4 and section 7.5 we describe the evaluation process with detailed analyses on which relations help improve retrieval performance. We also show via human evaluation that the images retrieved by the proposed CMCM are preferred over baseline by a huge margin. Our work provides insights into the ways that different modalities communicate and the role of coherence relations in capturing commonsense inferences in text and imagery.
7.2 Datasets

We study the efficacy of CMCM for image-retrieval by leveraging two image-text datasets CITE++ and Clue [13] that are annotated with image-text coherence relations. CITE++ is extended by us from CITE [195] adding 2242 image-text pairs annotated with coherence relations.

(a) Once they have baked remove them from the oven and sprinkle lightly with sugar. After you have dressed them allow them to cool for about 5 minutes and serve

(b) Seals fighting for a spot to sleep on the rocks

Figure 7.2: Example image-text pairs from CITE++ (a) and Clue (b) datasets. Image-text pair on the left has relations Expansion, Elaboration and Temporal while the one on the right has relations Action as Visible

CITE++ is an extended version of the CITE dataset which itself is a subset of a popular recipe dataset RecipeQA [203]. The RecipeQA dataset consists of multimodal recipes that contains textual instructions accompanied by one or more images. CITE leveraged recipes that have one-to-one correspondence between instruction and image, e.g. every instruction in the text has one image that visualizes it. Using Amazon Mechanical Turk, the authors obtained answers to 10 questions that help characterize the relationship between image and text. We choose the questions that are best suited to train CMCM as described in Table 7.1. The original dataset has 2057 image-text pairs annotated with True/False answers to these questions indicating presence/absence of the coherence relation. To perform a more comprehensive experiment, we collected 2242 more pairs using the same annotation
protocol, giving us a total of 4299 image-text pairs. The distribution of relations in the entire dataset is given in Table 7.1. Figure 7.2 [a] shows an example from CITE++ dataset.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Positive rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visible</td>
<td>0.674</td>
</tr>
<tr>
<td>Subjective</td>
<td>0.066</td>
</tr>
<tr>
<td>Action</td>
<td>0.157</td>
</tr>
<tr>
<td>Story</td>
<td>0.243</td>
</tr>
<tr>
<td>Meta</td>
<td>0.391</td>
</tr>
<tr>
<td>Irrelevant</td>
<td>0.087</td>
</tr>
</tbody>
</table>

Table 7.2: Coherence relations and their distribution in Clue dataset [13]

The Clue dataset [13] is constructed using the much larger Conceptual Captions dataset [204] which is primarily an image captioning dataset like COCO [12]. Clue annotated 7559 image-caption pairs with six coherence relations as shown in Table 7.2 to summarize the structural, logical and purposeful relationships between the contributions of texts and images. Example image-caption pair with coherence relations are shown in Figure 7.2 (b). The distribution of relations in the Clue dataset is given in Table 7.2.

### 7.3 Cross Modal Coherence Model

In order to train CMCM for text-to-image retrieval, we leverage a state-of-the-art image retrieval model for recipes proposed in [200]. We write \( S = [w_1, w_2, ..., w_m] \) for the input natural language text composed of \( m \) words. (In principle \( w_i \) could be words, phrases, sentences or any other semantic unit of text.) Similarly, we write \( I \) for the corresponding image. Then the objective of an image retrieval model is to find the model parameter \( \theta \), such that \( \theta = \arg \max_{\theta} \Pr_{\theta}(I \mid S) \).

#### 7.3.1 Image Encoder

The image encoder \( E_I \) is a Resnet-50 [168] pretrained model followed by a bottleneck layer similar to that used in [200]. The image is first resized to \( 224 \times 224 \), and then forwarded
through $E_I$ to get the image embedding $f_I \in \mathbb{R}^{300}$.

### 7.3.2 Text Encoder

The text encoder $E_S$ starts from a pretrained word2vec model that embeds each word into a 300 dimensional vector. The word2vec model is trained using Gensim [205]. The maximum length of the text sequence considered is 200 for CITE++ and 40 for Clue based on the longest sentences in the dataset. Then, the word embeddings are given as input to a Long Short Term Memory (LSTM) network to get each word representation. We next apply an attention mechanism [104] to the LSTM representations similar to [200], which learns the attention for each word in the text and helps the model attend to word cues that help with predicting the coherence relation as well as to retrieve the correct image. Finally a fully-connected layer is applied to encode the joined representation of all words $h$ into the shared domain.

The outputs of the text and image encoders are then used with a triplet objective using cosine similarity trained with hard negative mining [200]. Let $s(a, b) = a^T b / \sqrt{(a^T a)(b^T b)}$ measures the cosine similarity between two vectors $a$ and $b$, then the objective for the retrieval task per sample is given by Equation 7.1,

$$
trip(a, p, n) = s(a, p) - s(a, n) - \alpha,
$$

$$
L_{ret} = \min \{ 0, \ trip(f_S^+, f_I^+, f_I^-) \} + \min \{ 0, \ trip(f_I^+, f_S^+, f_S^-) \},
$$

where $L_{ret}$ is the retrieval loss, $f_S^+$ and $f_I^+$ are outputs of text and image encoder for a pair of text and image while $f_S^-$ is a text output that does not correspond to current image and $f_I^-$ is an image output that does not correspond to current text. The margin $\alpha$ is set to 0.3 by cross-validation.
7.3.3 Coherence Aware Module

Instead of relying only on the encoders, we also leverage coherence relations labelled by humans. We add a Coherence Aware Module that takes the normalized features from both text and image encoders as input and then passes them through a Multi-Layer Perceptron to predict the relations.

The dimension of the final linear layer is equal to the number of relations in the dataset when trained with all relations (i.e. multi-label classification) and 1 when trained with a single relation (i.e. single-label classification). We use Binary Cross Entropy (BCE) as the loss function and the objective of Coherence Aware Module for one sample is,

$$L_{cls} = \sum_c w_c (y_c \log(x_c) + (1 - y_c) \log(1 - x_c)),$$

(7.2)

where $x_c$ is the probability assigned to relation $c$ by the model while $y_c$ is the ground truth binary value. Since the relations are not equally distributed in the dataset, we balance the training of different relations by giving a weight $w_c$ for each relation that is reciprocal to its proportion in the dataset.

The model is thus trained in a multi-task setting where the coherence predictor is the auxiliary task. The final objective over the entire batch with batch size $N$ is given in Equation Equation 7.3,

$$L_{total} = \frac{1}{N} \sum_{n=1}^{N} (L_{ret}^n + \lambda_{cls} L_{cls}^n),$$

(7.3)

where $\lambda_{cls}$ is the weight associated with the coherence aware module and is chosen empirically as described later.
7.3.4 Selective Similarity Refinement

The performance of the retrieval model depends on the similarities between a query caption \( S \) and all possible images \( \{I_i\}, i \in [1, \ldots, N] \) (including the ground-truth image). We use cosine similarity (though any other valid similarity metric can be used) and notate the similarities as \( \{\theta_i = \text{cosine}(S, I_i)\} \).

Leveraging Confidence Score

We use the coherence prediction from Coherence Aware Module to refine the similarity between an image–text pair for retrieval during inference. Note that we do not know the coherence relation characterizing a ground truth image–text pair. However, a well trained Coherence Aware Module is expected to predict coherence for a ground truth image–text pair with high confidence. We define a confidence function for a query caption \( S \) and one possible image \( I_i \) as

\[
\eta_{i,c} = e^{\lambda|x_c-0.5|}, \quad (7.4)
\]

\[
\eta_i = \sum_{c} \eta_{i,c}, \quad (7.5)
\]

where \( x_c \) is defined in Equation 7.2, and \( \lambda \) is a hyperparameter decided by cross validation. Confidence function with different \( \lambda \) are shown in Figure 7.3. We can see that lower \( \lambda \) decreases the impact of the confidence function. We define \( \lambda = 0.13 \) for CITE++ and \( \lambda = 0.12 \) for Clue datasets empirically. The refined similarity is defined as,

\[
\bar{\theta}_i = \theta_i \cdot \eta_i \quad (7.6)
\]
Selective Refinement

Though confidence score helps, by itself the score is a weak indicator performing only slightly better than random. We hence limit the use of confidence score to difficult examples. We hypothesize that similarity between a correct image–text pair should on average be “α” larger than that of a wrong image–text pair based on the definition of loss Equation 7.1. In Figure 7.4, we verify this hypothesis by plotting the rank of ground truth image vs. the difference between the similarities of the top 2 retrieved images with the query caption. We observe that when the difference between the similarities of the top 2 images (Δ) is large enough (e.g. ≥ 0.2), the retrieval is always successful (e.g. ground truth image rank = 1). Based on this analysis, we select difficult query captions as those with Δ < T, where T is a hyperparameter chosen as 0.1 empirically. We use the refined similarity Equation 7.6 for ”difficult” examples during inference.
7.4 Experimental Setup

7.4.1 Evaluation Metrics

We evaluate the retrieval performance of all the models similar to [200] using the median retrieval rank (MedR) and the recall at K (R@K) metrics. The retrieval range is set to be 500. We believe that a model trained with coherence relations would achieve higher Recall and lower Median Rank when compared with an agnostic model. However, unlike datasets like MSCOCO where text descriptions are very specific to the paired image, both CITE++ and Clue have image-text pairs that exhibit complex relationships. Given the diverse nature of image-text relations (e.g. Subjective and Action), we believe that the above quantitative metrics that measure if the model retrieves the original image is stricter. Hence, we also perform comprehensive user study to evaluate the performance of the model.

**MedR** \(0 \leq \text{MedR} \leq 1\) is computed as the median rank of the true positive over all queries, a lower MedR suggests better performance.

**R@K** \(0 \leq \text{R@K} \leq 100\) computes the percentage of true positives recalled among the top-
K retrieved candidates, a higher score indicates better performance. Here we only report the results of retrieving image by using the caption as query.

### 7.4.2 Training Details

In our experiments, we evaluate the model and the coherence relations on CITE++ and Clue datasets independently. We split the CITE++ dataset as 3439/860 for training/testing while the Clue dataset as 6047/1512 for training/testing. 10% of the training data is used as validation. Further training and hyperparameter details are given in the appendix.

We train all the models using the Adam [206] optimizer with an initial learning rate of $10^{-4}$. All models are trained for about 20 epochs and the epoch with the lowest MedR on validation data is chosen as the best epoch for inference. Throughout training, the model sees the text-image pair and the corresponding coherence relation as input. During inference, we give the text as input and retrieve the closest image in the shared space. The *Coherence Aware Module* is not used during inference.

For all models, we repeat the testing experiments three times and report the standard deviation of the main metric. The word2vec model for CITE++ and Clue datasets are trained on the training set with *minimum frequency* = 1 and *window size* = 10, which outputs a $\mathbb{R}^{300}$ vector for all the words in the vocabulary. The vocabulary size is 8917 for CITE++, whereas for Clue it is 5612. Training the retrieval model takes about 20 minutes on CITE++ and 35 minutes on Clue datasets with an NVIDIA K80 GPU and all codes were written using PyTorch [207] GPU library.

*Hyperparameter Search*

We determine the best value for the weight $\lambda_{cls}$ for CAM empirically as shown in Figure 7.5 (Left), $\lambda_{cls} = 0.10$ achieves the best MedR. Another key parameter is the maximum sequence length for the LSTM module in the text encoder, smaller value may lose information in the sentence and larger value may bring extra burden to the model. We observe
Figure 7.5: On CITE++ dataset, different values for $\lambda_{cls}$ vs MedR to determine the best value for $\lambda_{cls}$ as 0.1 (Left); Text sequence length vs MedR to determine the best value for maximum sequence length as 200 (Right).

in Figure 7.5 (Right) that $\text{length} = 200$ achieves the best MedR for CITE++ dataset. We did similar experiments for Clue dataset and decide these value to be $\lambda_{cls} = 0.10$ and $\text{length} = 40$.

<table>
<thead>
<tr>
<th>Model</th>
<th>Coherence Aware Module</th>
<th>Relations</th>
<th>Attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>×</td>
<td>-</td>
<td>×</td>
</tr>
<tr>
<td>CMCA</td>
<td>×</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>CMCM-NoAttn</td>
<td>✓</td>
<td>All</td>
<td>×</td>
</tr>
<tr>
<td>CMCM</td>
<td>✓</td>
<td>All</td>
<td>✓</td>
</tr>
<tr>
<td>CMCM$_c$</td>
<td>✓</td>
<td>$c$</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 7.3: Description of the models used for comparison. -NoAttn means removing the attention module from the proposed model. ‘All’ relations indicate that the Coherence Aware Module is trained with all the relations in a multi-label multi-class setting. $c$ indicates only one relation is used with the Coherence Aware Module in a binary classification setting.

7.4.3 Comparative Evaluation.

For both the datasets, we train the proposed model and compare with various baselines as shown in Table 7.3.
Table 7.4: Quantitative comparison of the models trained and evaluated on Clue dataset.

<table>
<thead>
<tr>
<th></th>
<th>MedR↓</th>
<th>R@1↑</th>
<th>R@5↑</th>
<th>R@10↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>19.8±1.9</td>
<td>11.6</td>
<td>28.6</td>
<td>38.3</td>
</tr>
<tr>
<td>CMCA</td>
<td>19.3±2.0</td>
<td>13.2</td>
<td>30.6</td>
<td>41.0</td>
</tr>
<tr>
<td>CMCM-NoAttn</td>
<td>20.6±2.6</td>
<td>12.4</td>
<td>28.9</td>
<td>38.8</td>
</tr>
<tr>
<td>CMCM</td>
<td>18.7±1.6</td>
<td>13.8</td>
<td>31.6</td>
<td>40.6</td>
</tr>
<tr>
<td>CMCM_Visible</td>
<td>19.6±3.1</td>
<td>13.4</td>
<td>31.7</td>
<td>41.1</td>
</tr>
<tr>
<td>CMCM_Subjective</td>
<td>25.0±3.1</td>
<td>12.9</td>
<td>29.4</td>
<td>38.0</td>
</tr>
<tr>
<td>CMCM_Action</td>
<td>20.9±2.1</td>
<td>11.7</td>
<td>28.4</td>
<td>38.0</td>
</tr>
<tr>
<td>CMCM_Story</td>
<td>17.7±1.7</td>
<td>13.0</td>
<td>30.7</td>
<td>41.5</td>
</tr>
<tr>
<td>CMCM_Meta</td>
<td>19.2±1.5</td>
<td>13.1</td>
<td>31.0</td>
<td>40.7</td>
</tr>
<tr>
<td>CMCM_Irrelevant</td>
<td>20.3±1.9</td>
<td>12.6</td>
<td>31.1</td>
<td>40.9</td>
</tr>
</tbody>
</table>

7.5 Results and Discussion

7.5.1 CMCM vs CMCA

The results on CITE++ dataset are shown in Table 7.5. We can see from the results that having attention over the text clearly improves the retrieval performance. This can be attributed to the lengthy texts in CITE++ dataset. Moreover, we observe that CMCM, trained with one coherence relation improves the performance of the model by 1% R@1, 3% R@5 and 1% R@10 on an average for all relations. Coherence Aware Module increases the performance on retrieval compared with the baseline even without attention (cf. CMCM-NoAttn vs CMCA). However, having all the relations CMCM does not perform better than the variants with only one relation, which confirms the conjecture that not all relations contribute in increasing the performance of the retrieval model.

The results on the Clue dataset are given in Table 7.4. We observe that both the attention mechanism and the coherence-aware module improve the performance of the baselines, but the combination of the two does not improve the performance further. We believe this is because, recall metrics are stricter for both CITE++ and Clue datasets, since some captions can adequately be visualized by more than one image.

To evaluate the contribution of selective similarity refinement using confidence score,
Figure 7.6: Comparison MedR between baseline, CMCA and different CMCM variants; as well as the comparison between the same model with and without selective similarity refinement. Left: CITE++ dataset. Right: Clue dataset

we compare MedR based on $\theta_i$ and $\bar{\theta}_i$ of the same model in Figure 7.6. It can be seen that in CITE++ dataset, CMCM variants with attention and relation usually performs better than the baseline and CMCA (whose performance is highlighted by a dashed line). In Clue dataset, the situation is a little complicate because not all relations contribute positively. However, in most cases, model using selective similarity refinement (i.e. sim*conf) performs better than the same model without refinement (i.e. sim), which proves the effectiveness of the refinement technique. Note for the CMCM trained with ‘Irrelevante’ relation on Clue dataset (last two bars on Figure 7.6 right), applying the refinement severely degrade the performance, which demonstrates that irrelevant relation can not help in the retrieval task.

7.5.2 Human Evaluation

Both CITE++ and Clue have image-text pairs with complex coherence relations in contrast to datasets like MSCOCO that have predominantly just Visible relation. Hence, considering the ground truth as gold standard is not reasonable. Given the wide distribution of different relations in the datasets, the quantitative metrics (e.g. MedR and Recalls) are unreliable for the proposed setting. Therefore, we perform human evaluation where the top 1 retrieved
(a) **Action**

Horse grazing on a summer meadow in the forest outdoors.

(b) **Visible**

A vector illustration of a happy male golfer.

(c) **Temporal_{i>t}**

Finishing - Paint all the black parts except the door on the locomotive with gold food paint... add more details.

Figure 7.7: The ground truth image (Left) and the top 5 retrieved images by the CMCM and the CMCA models for two examples. The coherence relation (in blue) and caption are given above the images. The image-text pair in example (a) has *Action* relation while in example (b) has *Visible* relation. In example (a) the CMCM model leverages the *Action* coherence relation to retrieve images that depict some action in the top 5. Similarly in example (c) images retrieved by proposed our model with CAM retrieves images that depict the result of a process as given by the *Temporal_{i>t}* relation, whereas the agnostic model shows images that depict action in progress.
Table 7.5: Quantitative comparison in CITE++ dataset. The relations corresponding to each $Q_i$ are shown in Table 7.1. $\downarrow$ indicates that lower the better and $\uparrow$ indicates that higher the better.

<table>
<thead>
<tr>
<th></th>
<th>MedR\downarrow</th>
<th>R@1\uparrow</th>
<th>R@5\uparrow</th>
<th>R@10\uparrow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>10.0±3.7</td>
<td>45.7</td>
<td>48.4</td>
<td>50.6</td>
</tr>
<tr>
<td>CMCA</td>
<td>5.4±2.3</td>
<td>46.0</td>
<td>50.1</td>
<td>53.8</td>
</tr>
<tr>
<td>CMCM-NoAttn</td>
<td>6.6±2.5</td>
<td>46.0</td>
<td>49.5</td>
<td>52.0</td>
</tr>
<tr>
<td>CMCM</td>
<td>4.2±1.2</td>
<td>46.5</td>
<td>51.4</td>
<td>53.9</td>
</tr>
<tr>
<td>CMCM$Q_2$</td>
<td>4.7±2.0</td>
<td>46.4</td>
<td>50.6</td>
<td>53.4</td>
</tr>
<tr>
<td>CMCM$Q_3$</td>
<td>4.2±1.3</td>
<td>46.2</td>
<td>51.1</td>
<td>54.2</td>
</tr>
<tr>
<td>CMCM$Q_4$</td>
<td>4.2±1.3</td>
<td>46.2</td>
<td>51.2</td>
<td>54.2</td>
</tr>
<tr>
<td>CMCM$Q_5$</td>
<td><strong>3.7±1.3</strong></td>
<td><strong>46.6</strong></td>
<td><strong>51.5</strong></td>
<td><strong>54.4</strong></td>
</tr>
<tr>
<td>CMCM$Q_6$</td>
<td>4.6±1.4</td>
<td>45.9</td>
<td>50.8</td>
<td>53.4</td>
</tr>
<tr>
<td>CMCM$Q_7$</td>
<td>3.9±1.7</td>
<td><strong>46.9</strong></td>
<td>51.2</td>
<td>54.1</td>
</tr>
<tr>
<td>CMCM$Q_8$</td>
<td>5.0±1.7</td>
<td>46.4</td>
<td>50.8</td>
<td>53.8</td>
</tr>
</tbody>
</table>

Table 7.6: Human evaluation results. Values indicate the percentage of samples for which humans preferred the output of CMCM, CMCA, both or neither.

<table>
<thead>
<tr>
<th></th>
<th>CMCM</th>
<th>CMCA</th>
<th>Same</th>
<th>Neither</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMCM$V_{isible}$</td>
<td><strong>24%</strong></td>
<td>17%</td>
<td>46%</td>
<td>13%</td>
</tr>
<tr>
<td>CMCM$Subjective$</td>
<td><strong>53%</strong></td>
<td>10%</td>
<td>7%</td>
<td>30%</td>
</tr>
<tr>
<td>CMCM$Story$</td>
<td><strong>40%</strong></td>
<td>10%</td>
<td>33%</td>
<td>17%</td>
</tr>
<tr>
<td>CMCM$Meta$</td>
<td><strong>56%</strong></td>
<td>9%</td>
<td>25%</td>
<td>10%</td>
</tr>
<tr>
<td>CMCM$Q_7$</td>
<td><strong>43%</strong></td>
<td>17%</td>
<td>27%</td>
<td>13%</td>
</tr>
</tbody>
</table>

images by CMCA and CMCM models are shown for pairwise comparison.

We recruit 250 participants through Amazon Mechanical Turk. All subjects were US citizens, agreed to a consent form approved by anonymized institution review board, and were compensated at an estimated rate of USD 15 an hour. We showed subjects the caption, the top image retrieved by the coherence aware and the coherence agnostic model for five relations from both the datasets and asked them to choose one of the following options: (1) I prefer image A (which showed the result of CMCM) (2) I prefer image B (which showed the result of CMCA) (3) The images are exactly the same (4) Neither of the images is a good match for this text.
The results are shown in Table 7.6. It can be seen that the images retrieved by the proposed model are preferred by humans. More importantly, the difference in preference is quite significant in contrast to the quantitative metrics. We can also see that the difference in preference between CMCM and CMCA is higher when the relation is *Subjective* or *Story* when compared to regular captions (see *Visible*), indicating the importance of explicitly modeling coherence relations for cross-modal understanding. The results of the t-test shows that the differences observed in CMCM and CMCA category are all statistically significant ($p < 0.01, t > 14.1$). The results of the sensitivity power analysis shows that our experiment detects effect sizes as small as 0.17 with a power and significance level of 95%.

Figure 7.8: Attention weights for CMCM and CMCA models for example (a) and (b) in Figure 7.7.
<table>
<thead>
<tr>
<th></th>
<th>GT</th>
<th>CMCM</th>
<th>CMCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Q2</td>
<td>Fill 30 cupcake tins with batter, 3/4 full. Bake in oven for about 17 minutes or until toothpick comes out clean. Cool in pans for about ten minutes and then finish cooling on cooling rack.</td>
<td><img src="image1.jpg" alt="Image" /> <img src="image2.jpg" alt="Image" /> <img src="image3.jpg" alt="Image" /> <img src="image4.jpg" alt="Image" /> <img src="image5.jpg" alt="Image" /></td>
<td><img src="image6.jpg" alt="Image" /> <img src="image7.jpg" alt="Image" /> <img src="image8.jpg" alt="Image" /> <img src="image9.jpg" alt="Image" /> <img src="image10.jpg" alt="Image" /></td>
</tr>
<tr>
<td>(b) Q3</td>
<td>Introduce sugar in a stream - Being sure to add it in a slow stream. Whisk the sugar into the eggs espresso powder salt and vanilla mixture. Be aware that too much sugar at same time may cause batter to be grainy.</td>
<td><img src="image11.jpg" alt="Image" /> <img src="image12.jpg" alt="Image" /> <img src="image13.jpg" alt="Image" /> <img src="image14.jpg" alt="Image" /> <img src="image15.jpg" alt="Image" /></td>
<td><img src="image16.jpg" alt="Image" /> <img src="image17.jpg" alt="Image" /> <img src="image18.jpg" alt="Image" /> <img src="image19.jpg" alt="Image" /> <img src="image20.jpg" alt="Image" /></td>
</tr>
<tr>
<td>(c) Q4</td>
<td>Make sure ingredients are evenly distributed but not over mixed. 2. Bake at 35-40 min at 350 degrees. 3. A toothpick will come out clean</td>
<td><img src="image21.jpg" alt="Image" /> <img src="image22.jpg" alt="Image" /> <img src="image23.jpg" alt="Image" /> <img src="image24.jpg" alt="Image" /> <img src="image25.jpg" alt="Image" /></td>
<td><img src="image26.jpg" alt="Image" /> <img src="image27.jpg" alt="Image" /> <img src="image28.jpg" alt="Image" /> <img src="image29.jpg" alt="Image" /> <img src="image30.jpg" alt="Image" /></td>
</tr>
<tr>
<td>(d) Q5</td>
<td>The cabbage need to be diced.</td>
<td><img src="image31.jpg" alt="Image" /> <img src="image32.jpg" alt="Image" /> <img src="image33.jpg" alt="Image" /> <img src="image34.jpg" alt="Image" /> <img src="image35.jpg" alt="Image" /></td>
<td><img src="image36.jpg" alt="Image" /> <img src="image37.jpg" alt="Image" /> <img src="image38.jpg" alt="Image" /> <img src="image39.jpg" alt="Image" /> <img src="image40.jpg" alt="Image" /></td>
</tr>
<tr>
<td>(e) Q6</td>
<td>A little dab will do ya. Now you have a little dab (1-2 tbsp) of coffee in the sugar. Now we will make the paste. Place the coffee back on the burner to finish cooking.</td>
<td><img src="image41.jpg" alt="Image" /> <img src="image42.jpg" alt="Image" /> <img src="image43.jpg" alt="Image" /> <img src="image44.jpg" alt="Image" /> <img src="image45.jpg" alt="Image" /></td>
<td><img src="image46.jpg" alt="Image" /> <img src="image47.jpg" alt="Image" /> <img src="image48.jpg" alt="Image" /> <img src="image49.jpg" alt="Image" /> <img src="image50.jpg" alt="Image" /></td>
</tr>
<tr>
<td>(e) Q7</td>
<td>After a few hours your popsicle should now be completely frozen so take it out of the freezer. Now remove the popsicle from the cup. If you cannot run the cup under hot water for a few seconds and it should slip.</td>
<td><img src="image51.jpg" alt="Image" /> <img src="image52.jpg" alt="Image" /> <img src="image53.jpg" alt="Image" /> <img src="image54.jpg" alt="Image" /> <img src="image55.jpg" alt="Image" /></td>
<td><img src="image56.jpg" alt="Image" /> <img src="image57.jpg" alt="Image" /> <img src="image58.jpg" alt="Image" /> <img src="image59.jpg" alt="Image" /> <img src="image60.jpg" alt="Image" /></td>
</tr>
</tbody>
</table>

Figure 7.9: Example input text, ground truth image, ground truth image-text coherence relation and the top 4 retrieved images by the proposed CMCM algorithm and CMCA for comparison on the CITE++ dataset.
<table>
<thead>
<tr>
<th></th>
<th>GT</th>
<th>CMCM</th>
<th>CMCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Action</td>
<td>Young woman drinking coffee to go on the street.</td>
<td><img src="image1.jpg" alt="Images" /> <img src="image2.jpg" alt="Images" /> <img src="image3.jpg" alt="Images" /> <img src="image4.jpg" alt="Images" /> <img src="image5.jpg" alt="Images" /></td>
<td><img src="image1.jpg" alt="Images" /> <img src="image2.jpg" alt="Images" /> <img src="image3.jpg" alt="Images" /> <img src="image4.jpg" alt="Images" /> <img src="image5.jpg" alt="Images" /></td>
</tr>
<tr>
<td>(a) Story</td>
<td>The flower that blooms in adversity is the most rare and beautiful of all.</td>
<td><img src="image1.jpg" alt="Images" /> <img src="image2.jpg" alt="Images" /> <img src="image3.jpg" alt="Images" /> <img src="image4.jpg" alt="Images" /> <img src="image5.jpg" alt="Images" /></td>
<td><img src="image1.jpg" alt="Images" /> <img src="image2.jpg" alt="Images" /> <img src="image3.jpg" alt="Images" /> <img src="image4.jpg" alt="Images" /> <img src="image5.jpg" alt="Images" /></td>
</tr>
<tr>
<td>(a) Subjective</td>
<td>newly built small house next to the sea and the beach.</td>
<td><img src="image1.jpg" alt="Images" /> <img src="image2.jpg" alt="Images" /> <img src="image3.jpg" alt="Images" /> <img src="image4.jpg" alt="Images" /> <img src="image5.jpg" alt="Images" /></td>
<td><img src="image1.jpg" alt="Images" /> <img src="image2.jpg" alt="Images" /> <img src="image3.jpg" alt="Images" /> <img src="image4.jpg" alt="Images" /> <img src="image5.jpg" alt="Images" /></td>
</tr>
<tr>
<td>(a) Visible</td>
<td>Illustration of a magnifying glass over a blue background.</td>
<td><img src="image1.jpg" alt="Images" /> <img src="image2.jpg" alt="Images" /> <img src="image3.jpg" alt="Images" /> <img src="image4.jpg" alt="Images" /> <img src="image5.jpg" alt="Images" /></td>
<td><img src="image1.jpg" alt="Images" /> <img src="image2.jpg" alt="Images" /> <img src="image3.jpg" alt="Images" /> <img src="image4.jpg" alt="Images" /> <img src="image5.jpg" alt="Images" /></td>
</tr>
</tbody>
</table>

Figure 7.10: Example input text, ground truth image, ground truth image-text coherence relation and the top 4 retrieved images by the proposed CMCM algorithm and CMCA for comparison on the Clue dataset.

7.5.3 Qualitative Analysis

To further understand the behavior of the model, we investigate the attention weights over input text for CMCM and CMCA models. In example Figure 7.7 (a), the proposed coherence-aware model retrieves the ground truth within the top 5 images. We can see from Figure 7.8 (left) that adding Coherence Aware Module increases the weight on words horse and grazing relative to the agnostic model. This can be attributed to the model’s ability to predict the associated coherence relation to help retrieve the right image. The CMCA model, however, attends more to commonly visualized words like forest and outdoors.
Similarly, in Figure 7.7 (right), CMCM shows improved attention weights for words like male and golfer. The result is the model being able to retrieve the correct image in top 1 though both models retrieve images of Vector illustration in the top 5 Figure 7.7 [b]. More examples for other relations are provided in the Appendix.

In CITE++ dataset, we observe similar behavior as shown in Figure 7.7 [c]. The relation Temporal \(i\to t\) characterizes the temporal correlation between an image and text where the image visualizes the result of the process described in the corresponding text. These relations are difficult to implicitly understand as the text is no different from any other step in the recipe. Training with Coherence Aware Module that explicitly models temporal relation improves the performance of image retrieval. For example, we can see in Figure 7.7 that all top 5 images retrieved by the CMCM are images that visualize the result of a process, in contrast to CMCA model that shows images of the step being carried out as well.

CITE++

More example predictions by the proposed coherence aware and the agnostic model are given in Figure 7.9. We can see not only from the correctly retrieved images but in most cases, all the top 5 images are quite relevant to both the input text and the coherence relation. Note that the coherence relation is not shown to the model. Moreover, these are examples of when the ground truth image has the depicted relation as TRUE. Consequently in all the examples, the probability of presence of the depicted coherence relation is above 0.5. We also provide the average precision of coherence relation prediction by the auxiliary Coherence Aware Module in Table 7.7

Clue

We show more examples from the Clue dataset in Figure 7.10. The coherence aware property of the proposed model is clearly seen from the examples. For the agnostic model, we believe there exists an inherent bias towards the most frequent relation. In the Clue dataset
this is the Visible relation and hence most of the retrieved images by the agnostic model concentrates on words that can be visually grounded.

<table>
<thead>
<tr>
<th>CMCM-NoAttn</th>
<th>Visible</th>
<th>Subjective</th>
<th>Action</th>
<th>Story</th>
<th>Meta</th>
<th>Irrelevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMCM</td>
<td>0.82</td>
<td>0.24</td>
<td>0.47</td>
<td>0.51</td>
<td>0.71</td>
<td>0.36</td>
</tr>
<tr>
<td>CMCM</td>
<td>0.82</td>
<td>0.28</td>
<td>0.47</td>
<td>0.50</td>
<td>0.71</td>
<td>0.36</td>
</tr>
<tr>
<td>CMCM_Visible</td>
<td>0.83</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>CMCM_Subjective</td>
<td>–</td>
<td>0.18</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>CMCM_Action</td>
<td>–</td>
<td>–</td>
<td>0.52</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>CMCM_Story</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.55</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>CMCM_Meta</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.72</td>
<td>–</td>
</tr>
<tr>
<td>CMCM_Irrelevant</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.41</td>
</tr>
</tbody>
</table>

Table 7.7: Average Precision of Coherence relation prediction using probabilities from the Coherence Aware Module.

<table>
<thead>
<tr>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMCM</td>
<td>0.86</td>
<td>0.20</td>
<td>0.64</td>
<td>0.30</td>
<td>0.32</td>
<td>0.70</td>
</tr>
<tr>
<td>Q2</td>
<td>0.89</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Q3</td>
<td>–</td>
<td>0.58</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Q4</td>
<td>–</td>
<td>–</td>
<td>0.78</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Q5</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.61</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Q6</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.63</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Q7</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.80</td>
</tr>
<tr>
<td>Q8</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 7.8: Average Precision of Coherence relation prediction using probabilities from the Coherence Aware Module. All models are the proposed CMCM variants. For relation specific models, the prefix CMCM is removed due to space constraints.

7.5.4 Predicting Coherence Relations

To further understand the effect and importance of coherence relations, we analyze the model’s ability to predict the presence of a coherence relation given the ground truth image and text. For this, we use the models trained using the original objective in Equation 7.3. We provide the ground truth image and text as input and calculate the Average Precision (AP) of coherence relation prediction. The results are provided in Table 7.8 for CITE++ dataset. The results for Clue dataset are provided in the Appendix. We can see that in most
cases as expected, the CMCM algorithms perform reasonably well. We haven’t provided the results for CMCA models as they were not trained with coherence relations. These results are comparable to similar experiments performed in [13] though in their experiment, classification was the only objective. Interestingly Subjective relation has very low AP (cf. appendix) similar to retrieval performance but the proposed model obtains significance gain in performance in the human evaluation. This reinforces the unreasonableness of the Recall and Rank metrics for this task.

7.5.5 Limitations and Improvements

Automating the understanding and generation of multimodal discourse requires a joint understanding of co-occurring images and text. Our study shows the effectiveness of cross-modal coherence modeling for text-to-image retrieval tasks.

We observe that the existing Recall based quantitative metrics for text-to-image retrieval are harsher and fail to meaningfully evaluate retrieval systems especially when image-text pairs can be characterized by different coherent relations. Future work involves developing new transformer-based coherence-aware metrics that can better measure the performance of retrieval models. Additionally, human evaluation of such subjective tasks must be standardized with a single common evaluation criteria and platform.

Though our Cross-Modal Coherence Model performs better than Cross-Modal Coherence Agnostic Model model, it maybe useful to validate the robustness of this observation across other text–to–image retrieval techniques. One way to do that would be to add Coherence Aware Module to existing state–of–the–art text–to–image retrieval techniques and evaluate the change in performance. Moreover, the dataset currently is still small to perform large experiments or train large complex models. An important next step is to improve the dataset by getting additional annotations.
CHAPTER 8
CONCLUDING REMARKS AND FUTURE DIRECTIONS

In this dissertation, we explored and took the first steps towards developing AI systems that can create and comprehend multimodal stories. We believe that a system that can comprehend stories can comprehend essentially any kind of data [14]. Motivated by the need for studying multimodal stories, we made the following contributions:

1. **Datasets:**
   We highlight limitations such as lack of coherence and ignorance of the creative process associated with visual storytelling in chapter 4. We then propose two major datasets MMSI (chapter 5) and AESOP (chapter 6) to overcome these limitations. MMSI ensures that a well performing model must have been able to comprehend coherence while AESOP captures the creative process associated with storytelling and paves way for modelling the modalities together.

2. **Tasks:**
   Moving on from tasks like Visual Storytelling [42] that fail to demand story comprehension, we propose Story Illustration (chapter 4) followed by a generalized extension Many–to–Many Story Illustration in chapter 5. These tasks are closer to imitating the narrative comprehension of humans through visualization. In chapter 6 we propose foundational visual storytelling tasks that are truly multimodal by formulating creation as co–evolution of visual and textual concepts.

3. **Neural Methods:**
   We introduce explicit coherence modelling in chapter 4 to model coherence between sentences in a story for story illustration in an end–to–end trained hierarchical neural architecture. Through chapter 5 we postulate story illustration as machine transla-
tion treating visual sequence of images as a different language describing the same context as the input story. We observe state of the art performance and ability to generalize to multiple datasets and settings. In chapter 6 we model the co-evolution of visual and textual concepts using sequentiaon co-attention networks. We also propose a Coherence Aware Module in chapter 7 that improves the performance of text to image retrieval by being aware of different coherent relations that characterize image-text pairs.

8.1 Future Work

Visual Storytelling as a field of research in AI is new and growing. Through this dissertation, we believe that the field will gravitate towards datasets, tasks and neural methods that are more similar to human cognition and communication. Though our work brings us closer to better AI systems, there still are limitations and scope for improvement.

8.1.1 Need for Better Evaluation Metrics

Story comprehension is a highly intellectual, subjective and creative process. Existing metrics that are used to evaluate natural language generation or image generation cannot be used to evaluate generated stories. Through several experiments (section 4.4, section 6.5) we have shown that existing quantitative metrics are insufficient with a huge variation from human evaluation. We proposed the visual saliency metric in section 4.3, but it is far from perfect to reasonably evaluate story comprehension.

Quantitative evaluation metrics help with fast prototyping and to give a reasonable notion of performance of a trained model. A standard evaluation metric or platform ensures different models and advancements can reasonably be compared and evaluated. Towards this goal, we would like to develop a standardized Question Answer style human-in-the-loop evaluation platform [208] that can be used to evaluate models developed for story comprehension as a fully automatic metric seems far fetched for story comprehension.
8.1.2 Modelling Causality and Creativity

Our systems enable modelling creativity but fail to explicitly model them through effective neural methods. As a first step towards modelling creativity, we want to work on generative models that can leverage the compositional properties of the AESOP dataset (section 6.2), to be creative. More specifically, we would like to work towards a Variational approach that can generate reasonable and plausible story endings or story visualizations conditioned on part of the story or other relevant prompts. Moreover these generations could be controlled by other metadata such as the story’s genre or title. Our current approach to modelling the visual scene composition is based on autoregressive techniques while we would like to make the scene generation non-autoregressive possibly based on scene graphs [209] or set prediction techniques [210].

The proposed AESOP dataset is versatile and has numerous possibilities. One of the main reasons to study stories is that, they are natural examples of cause and effect. Causality drives narratives and effectively communication and comprehension. It is important to explicitly model causality [22]. To model causality, we could annotate the events in stories with causal explanations grounded on cognitive psychology theory [122]. This would help us develop models that can be trained to provide explanations for why certain events follow other events learning causality in the process. On the other hand, given the exceptional advancements in image captioning and understanding, we could also use descriptive captions as a form of intermediate representation between visual and textual parts of a multimodal story or even as explanations. A story would then be a more abstract and causal linguistic rendering of the captions. A direct extension of this research is the application of models trained on this data to movies and videos that also have audio signals that can provide information that is not present in the vision or language modalities.

Systems that can comprehend stories will have the capability to understand and reason about human discourse, and naturally interact with humans. By taking a bottom up approach to research, starting from learning simple abstract notions of visual and textual
cues, to modelling context, coherence, compositionality, causality, creativity and story comprehension, we could take Artificial Intelligence a step closer to having human-like conversations, intellect, comprehension and creative abilities. Such systems find numerous applications in the real-world [14]. We believe that future research geared along the directions set forth above will lead to AI systems that can comprehend the visual world around them and comfortably and reasonably communicate with humans.
REFERENCES


[52] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual


[73] G. Radevski, G. Collell, M.-F. Moens, and T. Tuytelaars, “Decoding language spatial relations to 2D spatial arrangements,” in Findings of the Association for Com-


on Empirical Methods for Natural Language Processing (EMNLP), 2016, pp. 925–931.


