AESOP: Abstract Encoding of Stories, Objects, and Pictures

Hareesh Ravi[1]* Kushal Kafle[2] Scott Cohen[2] Jonathan Brandt[2]
Mubbasir Kapadia[1]
[1]Rutgers University, [2]Adobe Research
[1]{hr268, mk1353}@cs.rutgers.edu, [2]{kkafle, scohen, jbrandt}@adobe.com

Abstract

Visual storytelling and story comprehension are uniquely human skills that play a central role in how we learn about and experience the world. Despite remarkable progress in recent years in synthesis of visual and textual content in isolation and learning effective joint visual-linguistic representations, existing systems still operate only at a superficial, factual level. With the goal of developing systems that are able to comprehend rich human-generated narratives, and co-create new stories, we introduce AESOP*: a new dataset that captures the creative process associated with visual storytelling. Visual panels are composed of clip-art objects with specific attributes enabling a broad range of creative expression. 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. We benchmark the proposed approach with human baselines and evaluate using comprehensive qualitative and quantitative metrics. Our results highlight key insights related to the dataset, modelling and evaluation of visual storytelling for future research in this promising field of study.*

1. Introduction

"Examples are the best precept" – Aesop, *The Two Crabs*

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 [52]. 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 [36]. Consequently, stories are crucial to mental development in humans. We postulate that machine intelligence requires comparable skills, particularly when interacting with people.

Though there have been some works on understanding and modelling natural language stories [55, 57, 68], there is limited work on aligning stories with the visual world [33, 34]. When there is no visual information available as part of a story, such as in novels, people still inherently visualize the events in real-time to disambiguate details and make inferences about the story [95] with ease. 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. Particularly, the text, and benchmarks associated with images in popular datasets such as MSCOCO [47], Flickr [87] and Visual Genome [42] focus on superficial factual descriptions, rather than a narrative. Moreover, the few existing visual story datasets [33, 30] lack coherence and diversity [64] that are key to a good story. Also, these datasets assume visual storytelling as a perceptive process rather than a creative process. For example, in [33], crowd workers wrote natural language stories given a sequence of images from photo albums. Such a process leads to superficial and disjoint stories [3, 34] 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 a third of the stories in [33], 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 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 im-

*A portion of this work was done during Hareesh Ravi's internship at Adobe Research.

Heartfelt Advice
(Themes: sad, drama, moral)

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 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.

age sequence. The converse task, to generate visual input that would match the given text story, is markedly absent in the literature. Though some recent works developed techniques for generating visual input from text [41, 73, 61] they still focus on factual information extraction rather than narrative understanding.

In this paper, 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 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 [94, 73], our dataset employs abstract visual scenes, with a broad set of choices for objects and attributes needed for visual storytelling. Examples of the wide range of stories created from this diverse, yet finite palette, appear in supplementary materials.

Current visual storytelling research has dealt with tasks such as storytelling, generation [30, 33] and illustration [64] or cloze tasks in [34] 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.

Our contributions are as follows:

(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.

(2) 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.We will make the dataset publicly available[2] to promote future research in this promising and challenging field of study.

2. Related Works

2.1. Vision and Language Integration

There has been extensive research in multimodal data understanding with large and comprehensive datasets [47, 87]. Modelling techniques are usually based on joint embedding space learning [44, 23], text to image retrieval/generation [82, 84, 78, 61, 45] or image to text generation/retrieval [86, 5, 29, 78, 39, 51, 37] tasks. Some recent works [50, 74, 16, 19] have proposed large multimodal pretraining networks based on the Transformers [77] architecture that obtain state of the art results on more than one specific image-text understanding task.

There has been increasing interest in modelling the subjective attributes of image–text representation learning by associating an emotion label [71], hashtags [18], personalization [22] and cross–modal coherence labels [2] with image-text pairs. Other works along the same line include [91, 85]. Images are composed by clip–art objects in [62, 73, 94, 41] where the aim is to model image–text relationship from the perspective of abstract visual reasoning.

All these works focus on factual information extraction from an image (or vice versa) using descriptive text whereas

[1] Reference to Aesop, the Greek Fabulist and Storyteller.

[2] https://github.com/Hareesh-Ravi/AESOP

human language and reasoning is more abstract and subjective. Moreover, visual scene understanding is portrayed as a static problem where text is used to describe a single static image whereas Cognitive and Neuroscience literature [60, 52, 27] suggests that visual perception is an abstract and temporal process.

Videos provide the necessary temporal information to model visual scene understanding. Datasets for video-text alignment span movies [65, 13] to instructional videos [92, 72] including large pretrained multimodal transformers [24, 72, 53, 93, 58] for joint representation learning. Previous works on video captioning include [80, 70, 79].

Though these works address the temporal aspect of visual scene understanding, text describes the video in a factual manner rather than imitating the abstract and subjective aspects of day-to-day human discourse. Moreover, these works primarily perform text generation conditioned on videos or retrieve videos as a whole conditioned on the text but do not model the co-evolution of visual and textual modalities for scene or story creation.

2.2. Stories

Text Only: Narrative understanding has been extensively studied in natural language [12, 67]. Some works focus on story understanding from the perspective of learning scripts [66] while [11, 8, 17, 59] perform unsupervised learning of event schemas and narrative chains from stories. Recent datasets such as ROCstories [55] have accelerated deep learning for story comprehension research via datasets, standardized cloze style tasks and metrics. Towards this, [89] proposed a hierarchical plot plus story generator while [32] use common sense knowledge base conceptNet for story comprehension. Similarly [76] proposed a scene graph approach to story generation while [56] extended the dataset to provide causal event annotation to study causality in stories. Other similar works include [14, 54, 4, 90]. 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.

Visual Storytelling: Visual Storytelling was introduced in [33] with the VIsual StoryTelling (VIST) dataset. It contains sequences of five images spanning an average of 7.9 hours obtained from flickr albums, aligned with one sentence describing each image forming a story. A CNN-RNN baseline was shown to create meaningful stories. Following that, many works [48, 88, 81, 83, 31] addressed visual storytelling with techniques ranging from learning a joint embedding space to adversarial reward-based algorithms. There has also been research on the more challenging story illustration task proposed in [64, 15]. Following this work, [46] proposed a GAN framework for story generation instead of illustration, evaluated on cartoon dataset [35]. Other similar works include [40] that aligns photo streams with text segments of a blog while [1] formulate the problem of sorting jumbled images and captions to form a coherent story on the VIST dataset. [9] propose a variational recurrent network for step wise illustration of cooking recipes.

It is shown in [64, 48, 3] that visual coherence in the sequence of images in VIST dataset is highly variant and sometimes non-existent considering the long average time span between events [33]. 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 model a general coherent narrative.

3. AESOP Dataset

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 [64, 3]. 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.

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 [21]. We seek to limit the complexity of the worlds our authors can create by simplifying the visual palette available to them. Inspired by [94, 41, 73, 62], we choose a clip–art based scene representation to depict the stories. As outlined in [94], 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).

3.1. Data Acquisition Setup

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 [6], which is, in turn, an extension of the tool in [94]. 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 [6], scenes in AESOP consist of 20 human characters with deformable limbs representing various ages, 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. All objects, scenes and other configurations for our final tool are given in detail in the supplementary materials. 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.

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. More data statistics are provided in the appendix.

3.2. AESOP Vs. Other datasets

We comprehensibly analyze AESOP to study how it overcomes the limitations of existing vision–language and visual storytelling datasets highlighted in Sec. 1 and Sec. 2.

Diversity: Verbs in text can be used to provide a notion of diversity in a dataset [3]. Compared to VIST [33], MSCOCO [47] and Flickr [87], the AESOP dataset shows more frequent use of verbs (Table 1). 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 [3], we use the the American National Corpus ANC [43] 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 [3] 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.

Dataset	Verb Freq.	Top 30	Non-Visible Verbs		Visible Verbs	
			Worry	Wonder	Sit	Walk
AESOP	0.198	0.589	556.0	93.5	1412	1110.1
VIST	0.017	0.669	9.8	2.3	130.9	64.3
MSCOCO	0.026	0.724	0.1	0.1	683.5	1991.5
Flickr	0.012	0.723	0.1	0.4	524.6	675.0
ANC	0.184	0.563	143.6	196.1	264.4	269.1

Table 1. 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.

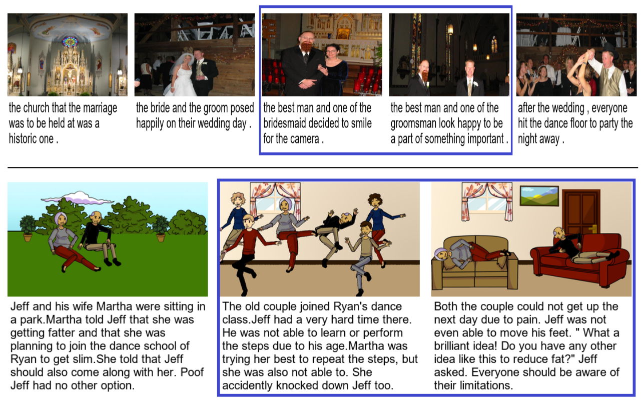

Figure 2. An example story from VIST (top) and AESOP (bottom) with two consecutive panels highlighted in Blue. Swapping the highlighted panels in VIST gives a story that is indistinguishable from the original showing lack of causality and coherence. In our dataset, swapping these panels would lead to a meaningless story.

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 2. It is hard to tell the correct order of panels in the VIST story, whereas in the AESOP example it is clear that *pain from the dance* is a result of the *dance* indicating clear causality.

4. Towards comprehension and co-creation 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 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:

4.1. 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 [34]. A human baseline example for assistant illustrator is shown in Figure 4. 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.

4.2. 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 [33]. A human baseline example for this task is shown in Figure 5. 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 [57, 34]. 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.

For both Assistant Illustrator and Writer, any of the three visual or text panels can be masked and predicted. However,

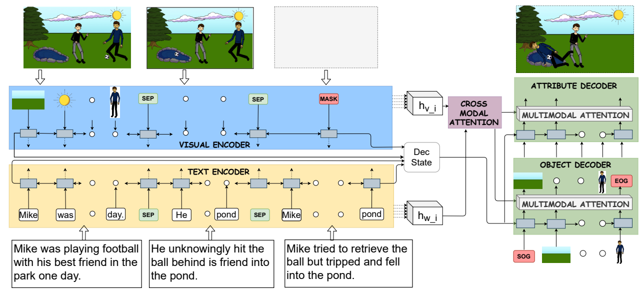

Figure 3. AESOP model architecture containing a Text and Panel Encoders, followed by cross–modal attention and hierarchical decoders to generate a visual panel. (Zoom in for details)

we limit our results, examples and analysis to completing the *final* missing panel for ease of presentation and comparison with human baselines. Results for arbitrary panel masking are present in the supplementary materials.

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. We discuss some possibilities in Sec 9.

5. AESOP Model

Following the approach of [41, 73, 62], we treat visual panels as a sequence of objects and attributes. Our overall model is shown in Figure 3.

5.1. 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 $\mathbf{V} = [v_0, v_1, v_2, ..., v_{nmax}]$ 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 [49] 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 [25] 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 [73, 62]. This ensures farthest objects like *sun*, *cloud*, *boat* are rendered first followed by other objects. Each object is then encoded as

$$\begin{aligned} what(v_i) &= LN(o_{emb}(o_i) + g(word(o_i))) \\ where(v_i) &= LN(f_{loc}([x_i; y_i; z_i; flip_i])) \\ how(v_i) &= LN(p_{emb}(pose_i) + e_{emb}(expr_i)) \\ f(v_i) &= what(v_i) + where(v_i) + how(v_i), \end{aligned} \tag{1}$$

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, 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 [62] but obtained better performance with this approach.

5.2. Story Encoder

Let [$\mathbf{V^1}$, $\mathbf{V^2}$, $\mathbf{V^3}$] be the sequence of visual panels that correspond to the sequence of text panels [$\mathbf{S^1}$, $\mathbf{S^2}$, $\mathbf{S^3}$]. Then we represent the entire story using sequences of visual and textual tokens as [$f(v_1^1)$, ..., $f(v_n^1)$, $f(v_1^2)$, ..., $f(v_n^2)$, $f(v_1^3)$, ..., $f(v_n^3)$] and [$g(w_1^1)$, ..., $g(w_n^1)$, $g(w_1^2)$, ..., $g(w_n^2)$, $g(w_1^3)$, ..., $g(w_n^3)$] respectively where $g(w_i^j)$ is the word embedding corresponding to the ith word in the jth 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 ⟨MASK⟩ token. Between the panels, we add a ⟨SEP⟩ token and in the beginning and the end ⟨SOS⟩ and ⟨EOS⟩ tokens respectively.

The story encoder consists of a visual, text and a cross–modal encoder. The visual and textual encoders are separate Bidirectional GRUs [7], 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 (more details in supplementary materials).

5.3. Panel Decoder

Visual Panel: We pose generation of the masked visual panel as [73] prediction of the following sequence $\mathbf{V} = [v_0, v_1, v_2, ..., v_n]$. 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 *where* attributes are clamped to be between 0 and 1 while *softmax* function is applied for pose and expression classification. Further details are provided in supplementary materials.

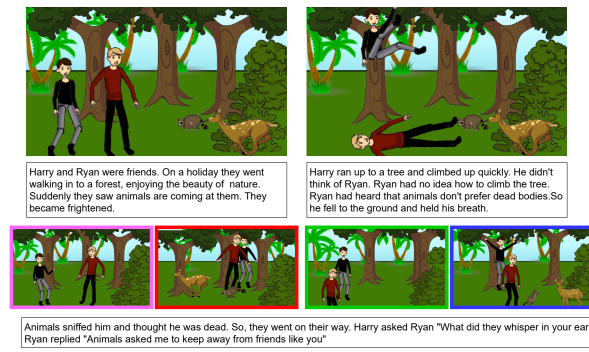

Figure 4. Examples of Assistant Illustrator result by Ground truth, Human Baseline, Proposed model and Unimodal are shown.

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. During inference, nucleus sampling [28] is used to generate the final text.

6. Baseline Models

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 decoder attention modules. For text, we fine-tune a pretrained GPT-2 model [63] on in–filling task [20], 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

Model↑	BG ↑	O-IOU ↑	Loc ↑	Dep ↑	Flip ↑	Pose ↑	Expr ↑	Scene↑	B–1↑	B–4 ↑	M ↑	R–L ↑	C ↑
	Illustrator								Writer				
Proposed	90.1	66.5	0.73	92.2	89.2	30.4	41.3	4.1	26.28	1.96	9.02	22.1	17.9
Unimodal	89.8	68.5	0.71	92.2	89.5	33.3	37.9	4.2	10.06	0.41	6.7	10.4	5.7
One-to-one	68.3	18.0	0.42	46.4	26.1	5.42	7.32	1.2	23.04	1.72	7.2	10.8	7.2
Pixel	52.3	15.7	0.25	21.5	10.6	3.54	5.12	1.0	8.62	0.73	5.4	7.98	4.32
Human	95	72.6	0.86	73.7	70.6	23.7	38.2	4.0	21.08	1.84	11.1	15.1	18.2
Repeat	90	79.3	0.91	94.6	90.1	36.9	37.8	5.1	–	–	–	–	–

Table 2. Results of all models on Assistant Illustrator and Assistant Writer modes. 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.

ResNet–18 [26] network and visual attention modules perform spatial attention similar to [73]. 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.

Figure 5. Examples of Assistant Writer result by Ground truth, Human Baseline, Proposed model and Unimodal are shown.

7. 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.

Assistant Illustrator: Following the works of [41, 73, 62] 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 [62]. Scene Similarity metric proposed in [41] 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 [41]).

Assistant Writer: For text generation, we use existing metrics BLEU-k, METEOR, CIDEr and ROUGE-L [69].

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 4). 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? **2) Relevant:** Is the generated content relevant to the corresponding content from alternate modality in the same panel?, and **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.

Experiment	Meaningful	Relevant	Coherent	Overall
Human	77.2	84.5	81.8	87.1
Proposed	6.6	7.1	7.0	7.6
No preference	16.2	8.4	11.2	5.3
Proposed	29.4	32	30.8	35.8
Unimodal	26.2	23.6	24	27.8
No preference	44.4	44.4	45.2	36.4
Human	68.8	85.2	80	86.5
Repeat	7.2	4.6	6.6	5.3
No preference	24	10.1	13.4	8.2

Table 3. Results of user study comparing models pairwise along three dimensions for Assistant Illlustrator. Values are given in % and *overall* indicates the overall preference between the two shown models.

8. Results

Assistant Illustrator: We see from Table 2 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 3. In contrast to the the observation in Table 2, we can see in the user study that human baseline clearly outperforms 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.

Assistant Writer: 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 (refer supplementary materials), 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.

9. Discussion

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.

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 2. Evaluation of vision and text models is already a tremendous challenge [38, 75, 10], which is made further difficult by creative aspects of storytelling in AESOP. We will plan to provide a platform to perform human evaluation using the defined dimensions in a standardized manner using Mturk to allow for a fair comparison with our baseline models.

Complexity of AESOP: Compared to closely related works such as [73, 62] for abstract scene generation, AESOP is highly complex. Tan et. al. in [73] 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.[73], 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 [73]: '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 [33, 57], where the requirements are either ill–posed [33] or framed as an easier retrieval setup [57].

Further possibilities with AESOP: 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.

Concluding Remarks: 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.

Acknowledgements: The research was supported in part by NSF awards IIS-1703883, IIS-1955404, and IIS-1955365. We thank the anonymous reviewers for their constructive feedback and the numerous Amazon Mechanical Turk workers for their contribution to this dataset.

References

[1] Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. Sort story: Sorting jumbled images and captions into stories. In *Proceedings, EMNLP*, pages 925–931, 2016. 3

[2] Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, 2020. 2

[3] Malihe Alikhani and Matthew Stone. “caption” as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67. ACL, June 2019. 1, 3, 4

[4] Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J Martin, and Mark O Riedl. Story realization: Expanding plot events into sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7375–7382, 2020. 3

[5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *proceedings, IEEE CVPR*, pages 6077–6086, 2018. 2

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015. 4

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*, 2015. 6

[8] Niranjan Balasubramanian, Stephen Soderland, Oren Etzioni, et al. Generating coherent event schemas at scale. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1721–1731, 2013. 3

[9] Vishwash Batra, Aparajita Haldar, Yulan He, Hakan Ferhatosmanoglu, George Vogiatzis, and Tanaya Guha. Variational recurrent sequence-to-sequence retrieval for stepwise illustration. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, pages 50–64, Cham, 2020. Springer International Publishing. 3

[10] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016. 8

[11] Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, 2009. 3

[12] Eugene Charniak. *Toward a model of children's story comprehension.* PhD thesis, Massachusetts Institute of Technology, 1972. 3

[13] David L Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics, 2011. 3

[14] Jiaao Chen, Jianshu Chen, and Zhou Yu. Incorporating structured commonsense knowledge in story completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6244–6251, 2019. 3

[15] Shizhe Chen, Bei Liu, Jianlong Fu, Ruihua Song, Qin Jin, Pingping Lin, Xiaoyu Qi, Chunting Wang, and Jin Zhou. Neural storyboard artist: Visualizing stories with coherent image sequences. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2236–2244, 2019. 3

[16] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020. 2

[17] Jackie Chi Kit Cheung, Hoifung Poon, and Lucy Vanderwende. Probabilistic frame induction. *NAACL HLT 2013*, pages 837–846, 2013. 3

[18] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 895–903, 2017. 2

[19] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020. 2

[20] Chris Donahue, Mina Lee, and Percy Liang. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, 2020. 6

[21] Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. To test machine comprehension, start by defining comprehension. In *Proceedings, ACL*, pages 7839–7859, Online, July 2020. Association for Computational Linguistics. 3

[22] Thibaut Durand. Learning user representations for open vocabulary image hashtag prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[23] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2121–2129. Curran Associates, Inc., 2013. 2

[24] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision (ECCV)*, volume 5. Springer, 2020. 3

[25] Oran Gafni and Lior Wolf. Wish you were here: Context-aware human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 6

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings, IEEE CVPR*, pages 770–778, 2016. 7

[27] Fritz Heider and Marianne Simmel. An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259, 1944. 3

[28] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. 6

[29] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6), Feb. 2019. 2

[30] Junjie Hu, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. What makes a good story? designing composite rewards for visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7969–7976, 2020. 1, 2

[31] Junjie Hu, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. What makes a good story? designing composite rewards for visual storytelling. In *proceedings, AAAI*, New York, USA, February 2020. 3

[32] Shanshan Huang, Kenny Q. Zhu, Qianzi Liao, Libin Shen, and Yinggong Zhao. Enhanced story representation by conceptnet for predicting story endings. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, CIKM '20, page 3277–3280, New York, NY, USA, 2020. Association for Computing Machinery. 3

[33] Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *NAACL*, 2016. 1, 2, 3, 4, 5, 8

[34] Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daume, and Larry S Davis. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *Proceedings, IEEE CVPR*, pages 7186–7195, 2017. 1, 2, 5

[35] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *proceedings, IEEE CVPR*, pages 2901–2910, 2017. 3

[36] Philip N. Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010. 1

[37] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20, 2017. 2

[38] Kushal Kafle, Robik Shrestha, and Christopher Kanan. Challenges and prospects in vision and language research. *Frontiers in Artificial Intelligence*, 2:28, 2019. 8

[39] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *proceedings, IEEE CVPR*, June 2019. 2

[40] G. Kim, Seungwhan Moon, and L. Sigal. Joint photo stream and blog post summarization and exploration. In *proceedings, IEEE CVPR*, pages 3081–3089, 2015. 3

[41] Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513, Florence, Italy, July 2019. Association for Computational Linguistics. 2, 3, 5, 7

[42] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 1

[43] Geoffrey Leech, Paul Rayson, et al. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge, 2014. 4

[44] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, October 2019. 2

[45] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *proceedings, IEEE CVPR*, June 2019. 2

[46] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *proceedings, IEEE CVPR*, June 2019. 3

[47] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 1, 2, 4

[48] Yu Liu, Jianlong Fu, Tao Mei, and Chang Wen Chen. Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. Proceedings, AAAI, February 2017. 3

[49] SP Lloyd. Least square quantization in pcm. bell telephone laboratories paper. published in journal much later: Lloyd, sp: Least squares quantization in pcm. *IEEE Trans. Inform. Theor.(1957/1982)*, 18, 1957. 5

[50] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 2

[51] Yadan Luo, Zi Huang, Zheng Zhang, Ziwei Wang, Jingjing Li, and Yang Yang. Curiosity-driven reinforcement learning for diverse visual paragraph generation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 2341–2350. Association for Computing Machinery, 2019. 2

[52] Raymond A. Mar and Keith Oatley. The function of fiction is the abstraction and simulation of social experience. *Perspectives on Psychological Science*, 3(3):173–192, 2008. PMID: 26158934. 1, 3

[53] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 3

[54] Yusuke Mori, Hiroaki Yamane, Yusuke Mukuta, and Tatsuya Harada. Toward a better story end: Collecting human evaluation with reasons. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 383–390, 2019. 3

[55] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June 2016. Association for Computational Linguistics. 1, 3

[56] Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online, Nov. 2020. Association for Computational Linguistics. 3

[57] Nasrin Mostafazadeh, Lucy Vanderwende, Wen-tau Yih, Pushmeet Kohli, and James Allen. Story cloze evaluator: Vector space representation evaluation by predicting what happens next. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 24–29, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. 1, 5, 8

[58] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *European Conference on Computer Vision*, pages 336–352. Springer, 2020. 3

[59] Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret, and Romaric Besançon. Generative event schema induction with entity disambiguation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 188–197, 2015. 3

[60] Keith Oatley and Nicola Yuill. Perception of personal and interpersonal action in a cartoon film. *British journal of social psychology*, 24(2):115–124, 1985. 3

[61] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *proceedings, IEEE CVPR*, June 2019. 2

[62] Gorjan Radevski, Guillem Collell, Marie-Francine Moens, and Tinne Tuytelaars. Decoding language spatial relations to 2D spatial arrangements. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4549–4560, Online, Nov. 2020. Association for Computational Linguistics. 2, 3, 5, 6, 7, 8

[63] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 6

[64] Hareesh Ravi, Lezi Wang, Carlos Muniz, Leonid Sigal, Dimitris Metaxas, and Mubbasir Kapadia. Show me a story: Towards coherent neural story illustration. In *proceedings, IEEE CVPR*, June 2018. 1, 2, 3

[65] Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh, and Bernt Schiele. Generating descriptions with grounded and co-referenced people. In *proceedings, IEEE CVPR*, 2017. 3

[66] Roger C Schank and Robert P Abelson. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press, 2013. 3

[67] Lenhart K Schubert and Chung Hee Hwang. Episodic logic meets little red riding hood: a comprehensive natural representation for language understanding. In *Natural language processing and knowledge representation: language for knowledge and knowledge for language*, pages 111–174. 2000. 3

[68] Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. Story cloze task: UW NLP system. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 52–55. Association for Computational Linguistics, Apr. 2017. 1

[69] Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799, 2017. 7

[70] Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. Weakly supervised dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1916–1924, 2017. 3

[71] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. Engaging image captioning via personality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12516–12526, 2019. 2

[72] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019. 3

[73] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating compositional scenes from textual descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6710–6719, 2019. 2, 3, 5, 6, 7, 8

[74] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods*

in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5103–5114, 2019. 2

[75] Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and A. V. Hengel. On the value of out-of-distribution testing: An example of goodhart's law. *ArXiv*, abs/2005.09241, 2020. 8

[76] Zhixing Tian, Yuanzhe Zhang, Kang Liu, Jun Zhao, Yantao Jia, and Zhicheng Sheng. Scene restoring for narrative machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3063–3073, 2020. 3

[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[78] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2016. 2

[79] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings, IEEE ICCV*, pages 4534–4542, 2015. 3

[80] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7622–7631, 2018. 3

[81] Jing Wang, Jianlong Fu, Jinhui Tang, Zechao Li, and Tao Mei. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *Proceedings, AAAI*, 2018. 3

[82] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *proceedings, IEEE CVPR*, pages 5005–5013, 2016. 2

[83] Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings, ACL*, pages 899–909, 2018. 3

[84] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *ICCV*, October 2019. 2

[85] Shuang Wu, Shaojing Fan, Zhiqi Shen, Mohan Kankanhalli, and Anthony KH Tung. Who you are decides how you tell. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4013–4022, 2020. 2

[86] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015. 2

[87] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1, 2, 4

[88] Licheng Yu, Mohit Bansal, and Tamara Berg. Hierarchically-attentive rnn for album summarization and storytelling. In *EMNLP*, 2017. 3

[89] Meng-Hsuan Yu, Juntao Li, Danyang Liu, Dongyan Zhao, Rui Yan, Bo Tang, and Haisong Zhang. Draft and edit: Automatic storytelling through multi-pass hierarchical conditional variational autoencoder. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(02):1741–1748, Apr. 2020. 3

[90] Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. Dcmn+: Dual co-matching network for multi-choice reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9563–9570, Apr. 2020. 3

[91] Wei Zhang, Yue Ying, Pan Lu, and Hongyuan Zha. Learning long-and short-term user literal-preference with multimodal hierarchical transformer network for personalized image caption. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9571–9578, 2020. 2

[92] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 3

[93] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8746–8755, 2020. 3

[94] C Lawrence Zitnick, Devi Parikh, and Lucy Vanderwende. Learning the visual interpretation of sentences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1681–1688, 2013. 2, 3, 4

[95] Rolf A Zwaan, Mark C Langston, and Arthur C Graesser. The construction of situation models in narrative comprehension: An event-indexing model. *Psychological science*, 6(5):292–297, 1995. 1