Text-Only Training for Visual Storytelling

Yuechen Wang[1],   Wengang Zhou[1,2†],   Zhenbo Lu[2†],   Houqiang Li[1,2]

[1]CAS Key Laboratory of Technology in GIPAS, EEIS Department,
University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
wyc9725@mail.ustc.edu.cn,zhwg@ustc.edu.cn,luzhenbo@iai.ustc.edu.cn,lihq@ustc.edu.cn

ABSTRACT

Visual storytelling aims to generate a narrative based on a sequence of images, necessitating both vision-language alignment and coherent story generation. Most existing solutions predominantly depend on paired image-text training data, which can be costly to collect and challenging to scale. To address this, we formulate visual storytelling as a visual-conditioned story generation problem and propose a text-only training method that separates the learning of cross-modality alignment and story generation. Our approach specifically leverages the cross-modality pre-trained CLIP model to integrate visual control into a story generator, trained exclusively on text data. Moreover, we devise a training-free visual condition planner that accounts for the temporal structure of the input image sequence while balancing global and local visual content. The distinctive advantage of requiring only text data for training enables our method to learn from external text story data, enhancing the generalization capability of visual storytelling. We conduct extensive experiments on the VIST benchmark, showcasing the effectiveness of our approach in both in-domain and cross-domain settings. Further evaluations on expression diversity and human assessment underscore the superiority of our method in terms of informativeness and robustness.

CCS CONCEPTS

• Computing methodologies → Natural language generation; Scene understanding.

KEYWORDS

Visual Storytelling, Text-Only Training, Story Planning

---

*Corresponding authors: Wengang Zhou and Zhenbo Lu.

Captions:
(a) A group of people waiting near a street.
(b) A group of veterans in uniform are walking in a parade.
(c) A parade with student wearing blue walking by.
(d) The color guard is showing their flag twirling prowess.
(e) Several people are riding in an older blue model convertible.

Story:
We all waited anxiously for the parade to start. The vfw led it off with the flags. Then came our high school marching band. The color guard did an awesome job. The grand marshal finished the parade up in a beautiful classic car.

Figure 1: Example difference between image captioning and visual storytelling. Words highlighted in red represents contents in the corresponding image, and words highlighted in green represents subjective expressions and information reasoned from other images.

1 INTRODUCTION

Visual storytelling [1], a task aimed at generating narratives based on image sequences, has received significant interest due to its potential applications in diverse domains such as advertising, entertainment, and education. In comparison to other vision-to-language generation tasks, such as visual captioning [2], visual storytelling presents unique challenges stemming from its subjective and imaginative nature. As illustrated in Fig. 1, to create a coherent story that aligns with the visual input, each sentence must not only describe the corresponding image, but also maintain logical connections to both preceding and subsequent sentences. This dual requirement of ensuring cross-modality consistency while preserving narrative coherence constitutes the primary challenge of visual storytelling.

Existing works often require large amounts of labeled data and attempt to learn both cross-modality alignment and story coherence simultaneously through end-to-end training. By training on large manually annotated data, these models are capable of generating coherent and visual-related stories. Subsequent advancements, such as the incorporation of external knowledge [3] and

scene graphs [4], have further enriched generated stories with additional details. More recently, the employment of large pre-trained Transformer-based language models has led to considerable improvements in visual storytelling [5]. Nevertheless, the substantial cost associated with annotating and training extensive datasets remains a significant bottleneck, limiting the scalability of visual storytelling approaches.

On the other hand, the burgeoning capabilities of pre-trained models offer potential for leveraging these models to transfer knowledge to downstream tasks such as visual storytelling, facilitating more data-efficient learning. To this end, some prior works have combined generative language models [6–8] with cross-modality pretrained models [9] to explore text-only training for image captioning [10, 11]. However, while these cross-modality models trained on paired image-text data successfully align text with individual images, they are limited in their capacity to comprehend the temporal structure of image sequences—an essential component of visual storytelling.

Motivated by the observations discussed above, we propose a novel framework that leverages pretrained generative language models and cross-modality models for data-efficient visual storytelling. We formulate visual storytelling as a visual-conditioned story generation task. As shown in Fig. 2, we first fine-tune a pretrained language model using only textual data to develop a story generator. Then, we incorporate visual clues during the generation process. Specifically, at each decoding step, we use a pretrained cross-modality model CLIP [9] as a visual discriminator to compute a matching score between candidate text and input images. A visual condition planner is then designed to aggregate the matching results of the input images, emphasizing semantics in the corresponding image while retaining information from other images. Finally, the aggregated result is incorporated into the decoding probability distribution to guide the generation of the next token, resulting in a coherent and visually aligned story.

To demonstrate the effectiveness of our proposed method, we conduct extensive experiments on the widely-used VIST benchmark [1]. The results show that our approach achieves state-of-the-art performance on various evaluation metrics including comparing-based automatic metrics, statistics-based metrics, and human evaluation. Additionally, our method exhibits impressive generalization ability in domain-transfer experiments, suggesting its potential for real-world applications.

We summarize the major contributions of this work as follows:

- We formulate visual storytelling as a visual-conditioned generation problem and propose a data-efficient framework which is trained solely on text-only data by leveraging pretrained CLIP model.
- We introduce a visual condition planner which is free of training. The planner aggregates sequential visual inputs to provide local details while maintaining the global theme of the image album, thereby improving the quality of generated stories.
- Extensive experiments on VIST benchmark demonstrate the effectiveness of our proposed method, as evidenced by its superior performance compared to existing methods in both automatic metrics and human evaluations.

2 RELATED WORK

The main idea of our work is to model visual storytelling as a controlled text generation task, and exploit large pretrained models to reduce the cost of cross-modality training. In this section, we provide a brief review of the related areas.

2.1 Visual Storytelling

Visual storytelling was first introduced by Huang et al.[1], which involves the use of a sequence of images to convey a narrative, necessitating reasoning over temporal context rather than merely understanding a static moment. Early approaches expanded upon conventional image captioning models by learning contextualized image representations[12] and incorporating global visual information [13]. Additionally, reinforcement learning was employed to learn an implicit reward function through adversarial reward learning, optimizing the policy model to better align with human demonstrations [14]. Hierarchical architectures [15] and hierarchical reinforced training [16] have also demonstrated effectiveness in learning high-level semantics.

Given the imaginative nature of storytelling, external knowledge graphs have been integrated to introduce fictional concepts not present in images [3, 17, 18]. To provide richer stories with greater visual detail, Wang et al.[4] incorporated scene graph generation, while Li et al.[19] learned cross-modal rules for mining visual concepts. Braude et al.[20] proposed an ordered image attention approach to enhance story coherence through consistent grounding across sequenced images. Furthermore, Transformer-based frameworks have demonstrated capabilities in modeling spatial relationships between objects in images[21]. In light of the proliferation of large pre-trained models, several studies have focused on leveraging pre-trained models (PTMs) for visual storytelling. Strategies include fine-tuning pre-trained Transformer encoders [22, 23] and jointly tuning pre-trained language generation models with pre-trained image encoders [5].

While the aforementioned approaches have demonstrated improvements in generated stories by incorporating external models, knowledge, and annotations, they also result in a significant increase in computational cost. In contrast, our proposed method circumvents the challenges associated with cross-modality training and annotation expenses by exclusively focusing on training using a text corpus.

2.2 Controlled Text Generation

In natural language generation, incorporating controllable constraints for open-ended text generation is both important and fundamental [24]. With the advancements in pretraining, recent efforts have primarily concentrated on adapting pre-trained language models (LMs) to various attributes. A straightforward approach involves fine-tuning a pre-trained LM to generate text with specific attributes [25–28]. Alternatively, it is feasible to design new large LM architectures or retrain large conditioned LMs from scratch [29–31]. Recently, the exponentially increasing scale and capacity of pre-trained LMs have made it more viable and promising to fix pretrained parameters and guide generation through post-processing. Dethathri et al.[32] first proposed this paradigm as Plug-and-Play language models, wherein an attribute discriminator updates LM

Figure 2: The training and inference pipeline of our method. During training, we only train the language generator on a story dataset without visual information. Then, at inference time, we utilize a pretrained CLIP model as a visual discriminator to align images with candidate tokens. Additionally, we introduce a visual condition planner that aggregates image sequences, and the output visual control is then incorporated into the generation process.

hidden states through back-propagation for attribute-controlled text generation. To reduce the computational cost associated with classifier-like discriminators ranking generated text, fine-tuned small LMs have been employed as generative discriminators to guide the generation of large pre-trained LMs[33–36]. Pascual *et al.* [37] extended the plug-and-play method to keyword constraints and designed a distribution shifting strategy to augment the decoding probability of keywords.

Guided decoding methods have demonstrated remarkable flexibility in accommodating various constraint types and hold considerable potential due to their independence from language models. In this work, we model visual storytelling as a visual-conditioned story generation task and propose a visual-linguistic discriminator to guide the generation process.

2.3 Large Pretrained Models

Generative language models. Taking advantage of the parallelism in the Transformer architecture [38], generative language models have shown a remarkable improvement in their capabilities in the past few years. These models can be broadly classified into two categories based on their network architecture: Decoder-Only models [6, 8] and Encoder-Decoder models [39, 40]. Pretrained on large corpora, these models can effectively transfer to various language generation tasks, such as summarization, question answering, and story generation, with limited or even no supervised data.

Cross-modality pretrained models. As the foundation of visual-language understanding, the idea to align the two modalities and learn a joint embedding space has been investigated extensively in the past decade [41–44]. In recent years, large cross-modality aligning models based on Transformers have gained considerable

attention [9, 45, 46]. A representative work is CLIP [9], which trains two encoders for image and text inputs using a contrastive loss. With 400 million data pairs for training, CLIP has demonstrated remarkable zero-shot capabilities on multiple downstream tasks.

3 PRELIMINARIES

A standard generative language model predicts the probability distribution of the next token based on previous inputs, which can be formulated as $P_{LM}(x_t|x_{<t})$. As a result, the probability of a text sequence $\boldsymbol{x} = \{x_1, \ldots, x_T\}$ can be modeled as follows:

$$P_{LM}(\boldsymbol{x}) = \Pi_{t=1}^T P_{LM}(x_t|x_{<t}). \tag{1}$$

In order to incorporate controls during the generation process, a constraint c can be added to form a conditioned language model. This model generates the probability distribution of the next token based on the history inputs and the control constraint, and can be formulated as:

$$P(\boldsymbol{x}|c) = \Pi_{t=1}^T P(x_t|x_{<t}, c). \tag{2}$$

Krause et al. [33] designed a generative discriminator to predict the probability that every candidate text sequence corresponds to the given constraint, which is given as:

$$P_\theta(c|x_t, x_{<t}) = \frac{P(c)\Pi_{t=1}^T P(x_t|x_{<t}, c)}{\sum_{c' \in \{c, \bar{c}\}} P(c')\Pi_{t=1}^T P(x_t|x_{<t}, c')} \tag{3}$$

where θ represents the learned parameters of the discriminator. Then, based on the Bayes rule, the conditioned language model can be decoupled as:

$$P(x_t|x_{<t}, c) \propto P_{LM}(x_t|x_{<t})P_\theta(c|x_t, x_{<t}). \tag{4}$$

Therefore, each step of the generation process is implemented by combining an unconditioned language modeling $P_{LM}(x_t|x_{<t})$, and

an attribute discriminator $P_\theta(c|x_t, x_{<t})$ with the guided decoding strategy as described in Eq. (4). Here the discriminator is trained externally and can be easily used with any language generator in a plug-and-play manner.

4 METHOD

Given a sequence of images $\mathcal{I} = \{I_1, \dots, I_N\}$, a visual storytelling approach aims to generate a multi-sentence story x by predicting the probability $P(x|\mathcal{I})$. To achieve this, we propose a framework that combines a text-only trained language generator, a pretrained visual discriminator, and a visual condition planner. Fig. 2 illustrates the training and inference pipeline of our method. During the training phase, we fine-tune the language generator using story text, while in the inference phase, we employ the pre-trained visual discriminator and the visual condition planner to guide the generation process.

4.1 Text-Only Training

Compared to other supervised visual storytelling methods, our approach offers a notable advantage in that it requires training only on a text corpus, resulting in significant cost reductions in both training and annotation efforts. Specifically, we fine-tune a Transformer decoder-based language model on a text story corpus to bridge the gap between pretraining on generic text and generating coherent stories. Given a narrative text sequence $x = \{x_1, \dots, x_T\}$, the language model is fine-tuned by minimizing the maximum likelihood estimation (MLE) loss:

$$\mathcal{L}_{MLE} = -\frac{1}{T} \sum_{t=1}^{T} \log P_{LM}(x_t|x_{<t}) \quad (5)$$

Inspired by Su et al. [47], we incorporate an additional contrastive objective $\mathcal{L}_{CL}$ to encourage the generation of diverse and distinct expressions. The objective is defined as:

$$\mathcal{L}_{CL} = \frac{1}{T(T-1)} \sum_{i=1}^{T} \sum_{j=1, j\neq i}^{T} \max(0, \epsilon - s(x_i, x_i) + s(x_i, x_j)), \quad (6)$$

where ϵ is a predefined margin, and s is the cosine similarity between tokens, defined by:

$$s(x_i, x_j) = \frac{h_{x_i}^T h_{x_j}}{|h_{x_i}||h_{x_j}|}. \quad (7)$$

The overall training objective of the language generator is the combination of the above two losses:

$$\mathcal{L} = \mathcal{L}_{MLE} + \alpha \mathcal{L}_{CL}, \quad (8)$$

where α is a hyper-parameter to balance the loss items.

After fine-tuning on a text story corpus, the language generator is able to generate coherent stories in a style that is aligned with the training data. However, since the generation process of the language generator is solely based on textual input, it may not take into account any visual content or the desired topic of the story. To address this, we introduce a visual discriminator and a visual condition planner to control the story topic and add details to the generated sentences.

Figure 3: Illustration of the visual condition planner.

4.2 Visual Discriminator and Story Planning

As previously mentioned, we consider visual storytelling as a visual-conditioned story generation task, and employ the guided decoding paradigm to integrate visual controls into the language generator. To achieve this, we introduce a visual discriminator and a visual condition planner to score candidate sequences during generation. The visual discriminator is implemented using a pretrained visual-linguistic aligning model, while the visual condition planner is a training-free weighting model which aggregates the text matching results of different images.

During each generation step t, our language generator predicts a probability distribution $P_{LM}(x_t|x_{<t})$ over the vocabulary V of possible next tokens, based on the context $x_{<t}$. To guide the selection of candidate tokens, we employ a pretrained CLIP [9] model as a visual discriminator $\mathbf{D}$. Although the CLIP model has been pretrained on a large-scale dataset of paired visual and textual data, the pretraining process does not specifically involve annotations for visual storytelling. Therefore, by utilizing the pretrained CLIP, our method does not require cross-modality training and is capable of handling open-domain visual input. This makes our approach data-efficient and more scalable than previous methods.

Specifically, we feed each candidate token x_t into the text encoder of CLIP along with the context tokens $x_{<t}$ to obtain a textual representation $f_{x_{1:t}}$. For each image I_j in the input album, we extract a visual representation f_{I_j} using the visual encoder of CLIP, where $j \in 1, \dots, N$. Then, the cosine similarity of $f_{x_{1:t}}$ and f_{I_j} is computed as:

$$\mathbf{D}(x_{1:t}, I_j) = \frac{f_{x_{1:t}} f_{I_j}}{|f_{x_{1:t}}||f_{I_j}|}, j \in \{1, \dots, N\}. \quad (9)$$

As the CLIP model is trained to map visual and textual input representations into a sharing space, the matching score $\mathbf{D}(x_{1:t}, I_j)$ measures the relevance between candidate sequence $x_{1:t}$ and the input image I_j.

Method	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE_L	CIDEr
Fully-Supervised Methods							
INet [48]	35.6	64.4	40.1	23.9	14.7	29.7	10.0
TAPM [5]	37.2	-	-	-	-	33.1	13.8
OIAVist [20]	36.8	68.4	42.7	25.2	15.3	30.2	10.1
KAGS [18]	36.2	70.1	43.5	25.2	14.7	31.4	11.3
Text-Only Trained							
Top-k	20.3	40.0	15.6	5.6	2.2	15.7	0.6
Nucleus	19.6	38.6	14.2	4.9	1.9	15.5	0.5
MAGIC	20.3	41.2	16.1	5.9	2.8	16.0	**1.3**
Ours	**23.0**	**43.7**	**20.2**	**9.2**	**4.5**	**17.3**	1.2

Table 1: Comparison with existing methods on VIST test set. "Fully-Supervised" methods are trained on paired data, "Text-Only Trained" methods are trained on the textual stories of VIST. The best results under each metric are highlighted in bold.

Method	ROCStories							WritingPrompts						
	M	B-1	B-2	B-3	B-4	R_L	C	M	B-1	B-2	B-3	B-4	R_L	C
Top-k	15.3	28.6	9.5	2.5	0.7	12.1	0.2	15.0	26.8	8.0	2.0	0.4	12.2	**0.2**
Nucleus	15.0	28.4	9.0	2.4	0.7	12.0	0.3	14.3	25.6	7.3	1.5	0.3	11.9	**0.2**
MAGIC	16.4	**29.7**	10.1	2.7	0.7	12.6	0.1	15.4	27.8	9.6	**2.9**	0.5	12.8	**0.2**
Ours	**16.6**	28.6	**11.5**	**3.8**	**1.2**	**12.9**	0.2	**16.2**	**28.8**	**9.9**	2.9	**0.9**	**13.7**	**0.2**

Table 2: Domain transfer results of text-only trained methods. The best results under each metric are highlighted in bold.

To ensure that the generated story aligns with the visual input fine-level semantics of the corresponding image to the sentence being generated and maintains the overall theme, we propose a visual condition planner. It aggregates the scores of the input images to derive a visual control for the current decoding step. Inspired by the work of Lin and Riedl [36], the planner does not require any training and achieves both global and local alignment through weighting and multiplication operations.

As depicted in Fig. 3, the visual condition planner computes a control weight for each input image based on the position of current sentence in the story. More precisely, the weight ω_j for image I_j is:

$$\omega_j = C \exp(-\frac{(i-j)^2}{2\sigma^2}), \qquad (10)$$

where $i \in \{1, \ldots, N\}$ represents the position of current sentence in the story, and C is a constant to normalize the weights and insure $\sum_{j=1}^{N} \omega_j = 1$. When $i = j$, the current sentence should be the exact description of image I_j, while remaining coherent to other images $I_{k \neq j}$. Therefore, the weight of I_j is the largest, and weight of I_k descends as the distance $|k - j|$ grows. Finally, the planner applies weighted multiplication on the scores of different images to obtain a unified matching score between the candidate sequence $x_{1:t}$ and the input images $\mathcal{I}$. Formally,

$$P_w(\mathcal{I}|x_t, x_{<t}) = \Pi_{j=1}^{N} \mathbf{D}(x_{1:t}, I_j)^{\omega_j}. \qquad (11)$$

It is worth noting that in our experiments, the aforementioned process is applied to a subset of the entire vocabulary, thereby reducing the computational cost of encoding and aligning candidate text. Specifically, we select the top K tokens predicted by the language generator as the subset $V_{(t)}^K$. Moreover, to eliminate the bias

of the cross-modality alignment results, we normalize the scores among candidate tokens. The final output of the visual condition planner can be written as:

$$P(\mathcal{I}|x_t, x_{<t}) = \frac{e^{(P_w(\mathcal{I}|x_t, x_{<t}))}}{\sum_{x_i \in V_{(t)}^K} e^{(P_w(\mathcal{I}|x_i, x_{<t}))}}. \qquad (12)$$

4.3 Visual-Conditioned generation

Given the token probability predicted by the language generator and the aggregated cross-modality matching score, visual storytelling can be decoupled into the combination of language modeling and cross-modality aligning. Similar to Eq. (4), the probability of next token x_t can be decoupled as follows:

$$P(x_t|x_{<t}, \mathcal{I}) \propto P_{LM}(x_t|x_{<t})P(\mathcal{I}|x_t, x_{<t})^{\gamma}, \qquad (13)$$

where the hyper-parameter γ controls the weight of visual information in the language generation process. While a higher value of γ can improve the alignment of visual semantics, it may also adversely affect the quality of the generated language. Finding the right balance between language and visual information is crucial for achieving high-quality visual storytelling.

Furthermore, inspired by Su et al. [49], we incorporate a degeneration penalty into Eq. (13) to prevent the repetitive degeneration problem. The final probability of visual-conditioned is formulated as:

$$P(x_t|x_{<t}, \mathcal{I}) = P_{LM}(x_t|x_{<t})P(\mathcal{I}|x_t, x_{<t})^{\gamma} - \\ \beta(\max(s(x_t, x_j), j \in \{1, \ldots, t-1\})), \qquad (14)$$

where β is a hyper-parameter to control the degeneration penalty strength, and $s(x_i, x_j)$ is defined in Eq. (7).

Yuechen Wang, Wengang Zhou, Zhenbo Lu, & Houqiang Li

Method	VIST-Text		ROC		WP	
	D-1	D-2	D-1	D-2	D-1	D-2
MAGIC	2.4	6.6	0.8	1.3	1.1	2.5
Ours	**4.7**	**17.2**	**5.5**	**18.8**	**7.4**	**26.5**

Table 3: Diversity evaluation results. "VIST-Text", "ROC", "WP" represents the training of language generator is conducted on the text part of VIST dataset, ROCStories, and WritingPrompts, respectively. "D-n" refers to "Distinct-n". The best results for each metric are highlighted in bold.

5 EXPERIMENTS

Dataset. We make evaluation on the widely-used VIST benchmark [1] for visual storytelling. VIST contains 210,819 images from 10,117 Flickr albums. Each sample in VIST contains five images selected from an album, and a five-sentence story is annotated as ground truth. After excluding broken images, the dataset contains 40,098, 4988, and 5050 samples for training, validation, and testing respectively. We use the test split of VIST as the evaluation benchmark in all experiments. Following previous works [5, 14], we evaluate at the album level, generating one story for each album regardless of different selected images. During the training stage, we use the text part of the VIST training split, where all names are replaced with special placeholders.

Implementation Details. The language generator is initialized with a pre-trained GPT-2 model, and fine-tuning is performed on 2 GTX3090 GPUs for 40,000 steps with batch size of 256. We set the training loss weight α to 1. To implement the visual discriminator, we ultilize a pretrained CLIP with ViT-base architecture as the image encoder. The visual-conditioned generation is performed on 1 GTX3090 GPU. In the reported results, we set the hyper-parameters K, γ, and β to 45, 1, and 0.01, respectively.

Evaluation Metrics. Following the existing works on the VIST benchmark, we adopt a set of automatic evaluation metrics including METEOR (M) [50], BLEU (B-n) [51], ROUGE_L (R_L) [52] and CIDEr (C) [53]. METEOR measures the semantic alignment between generated and reference sentences by leveraging WordNet. BLEU computes the unigram and n-gram overlap between generated and candidate sentences. ROUGE_L measures sentence-level similarity by computing the length of longest common subsequence. CIDEr evaluates the consensus based on n-grams and weights n-grams using Term Frequency Inverse Document Frequency (TF-IDF) to emphasize informative content. However, we note that these metrics, as they rely on word correspondence with the ground truth, may not fully capture the quality of open-ended generation tasks such as storytelling.

5.1 Quantitative Results

Comparison with Existing Methods. We compare the generation quality of our method with a text-only trained methods. First, we adopt top-k sampling [54] ($k = 40$) and nucleus sampling [55] ($p = 0.95$). Since these sampling-based decoding strategy takes no account of visual inputs, we consider them as the lower bound of the text-only trained methods. We also include MAGIC [49], which was proposed for image captioning and image-based story generation.

Figure 4: Human evaluation results. "Tie" means the annotator cannot choose the better story.

Method	M	B-1	B-2	B-3	B-4	R_L	C
Ours-Max	22.6	41.8	18.9	8.5	4.1	17.0	0.9
Ours-Mean	22.8	43.2	20.0	9.1	4.4	17.2	**1.3**
Ours-Local	22.4	42.1	19.4	8.7	4.2	17.2	1.0
Ours-Planner	**23.0**	**43.7**	**20.2**	**9.2**	**4.5**	**17.3**	1.2

Table 4: Evaluation results of different image album aggregation strategies. The best results for each metric are highlighted in bold.

MAGIC takes an image as input and generate text outputs by adding CLIP similarity scores on language model predicted probabilities. To extend MAGIC to the visual storytelling task, we average the representation of input image sequence to form the visual input of MAGIC. To provide a comprehensive comparison, we also report results of several fully-supervised baselines. INet [48], TAPM [5], OIAVist [20], and KAGS [18]

In Table 1, we present the comparison of our proposed method with existing fully-supervised and text-only trained methods. As expected, the fully-supervised methods trained on cross-modality paired data exhibit better performance compared to the text-only trained methods. However, our proposed method outperforms the text-only trained baselines on almost all metrics by a considerable margin, demonstrating the effectiveness of our visual-conditioned generation strategy.

Cross-domain Transfer. In order to evaluate the generalization ability of our method, we also explore cross-domain transfer by using story datasets of different domains in the text-only training stage. Specifically, we use **ROCStories** [56] and **WritingPrompts** [54] for training. The training split of ROCStories dataset contains 51,165 five-sentence commonsense stories. And the training split of WritingPrompts dataset contains 272,600 stories collected from Reddit's WRITINGPROMPTS forum[1]. The average length of WritingPrompts stories is 734.5, and the average number of sentences is 39.4, making it significantly larger than the VIST dataset and introducing a larger domain gap. During training, we exclude the story title and writing prompts to align with the VIST evaluation process.

[1]www.reddit.com/r/WritingPrompts/

Figure 5: Analysis of the effect of number of candidates K.

Figure 6: Analysis of the effect of control weight γ.

In Table 2, we compare the cross-domain transfer ability between our method and the text-only trained baselines. We observe a considerable drop in performance for all methods when evaluated on datasets from different domains. This is expected since the style, theme and topic of the stories are different across datasets. However, our method still outperforms others on most evaluation metrics, demonstrating its superior generalization ability.

Diversity Evaluation. To further assess the expressive diversity of the generated stories, we use Distinct-n which calculates the number of distinct n-grams of all generated stories [57]. The value is divided by the total number of generated tokens to avoid favoring long sentences. The results presented in Table 3 demonstrate that our method significantly outperforms the baseline in terms of diversity. This can be attributed to the ability of our method to attend to both global and local visual input, which results in more informative and diverse expressions. Additionally, it can be observed that the diversity of generated stories is relevant to the training corpus, which suggests that the incorporation of external text corpus can benefit visual storytelling.

Human Evaluation. As illustrated in previous works [14, 20], automatic evaluation metrics are insufficient for visual storytelling due to its subjective and imaginative nature. To obtain more reliable estimates, we also perform human evaluation. Following common practice, we randomly selected 150 examples from the test set, and invited 5 human annotators to rank generation results of different methods. Specifically, the annotators were asked to evaluate the stories based on three criteria: relevance, expressiveness, and concreteness. Relevance refers to whether the story covers the topic and main objects in the images. Expressiveness refers to whether the story is coherent, grammatically and semantically correct, and free of repetition. Concreteness refers to whether the story is narrative and concrete.

Fig. 4 shows the evaluation results of 5 human annotators. Our method outperforms MAGIC by a large margin in all three aspects. The dominance of our method is most significant in terms of Concreteness, indicating a greater ability to incorporate visual details in the generated stories. The expressiveness of MAGIC is better than the two other aspects, which reflects the fact that the language quality of our method is slightly affected by introducing fine-level visual control. Additionally, the "Tie" option is selected in a large

percentage in all three criteria, which has not been reported in previous methods [5, 18, 20, 48]. We believe the reason is that the overall quality of stories generated by text-only trained methods is lower than full-supervised methods, making it difficult to rank for human annotators.

5.2 Ablation Study

Impact of Visual Condition Planner. We conduct ablation experiments to analyze the effect of the visual condition planner, which aggregates the cross-modality matching result of input images. Specifically, we replace the aggregation process to three straight-forward strategies: 1) choosing the maximum matching score in all images, 2) averaging the scores of all images, and 3) using the score of the corresponding image. The evaluation results in Table 4 indicate that both strategies of viewing the images equally within the album ("Ours-Max" and "Ours-Mean") and focusing solely on the corresponding local image ("Ours-Local") have a negative impact on the quality of the generated stories.

Impact of Hyper-parameters. During the visual-conditioned generation, the selection of top-K candidate tokens to compute cross-modality matching score with visual inputs and the addition of visual control to the decoding process with a control weight γ in Eq. (14) are governed by predefined hyper-parameters. Therefore, it is important to analyze the influence of these hyper-parameters on the quality of generated stories.

From the results in Fig. 5, we observe that the performance improves with K when $K < 30$, and remains relatively stable when K continues to increase. However, when K is too large (> 60), the performance slightly decreases as K keep increasing. It is also worth noting that the inference time significantly increases as K increases. Therefore, we choose $K = 45$ in our experiments as it strikes a balance between performance and efficiency. The results in Fig. 6 demonstrate the significant impact of the control weight γ on the generation process. Specifically, when the control weight is too small, the generated stories tend to be disconnected from the visual input, while an excessively large control weight will lead to a disruption in the decoding process, thus deteriorating the overall quality of the generated text. These experimental findings align with our initial intuition and suggest the importance of selecting an appropriate control weight in the visual-conditioned generation.

GT	halloween night. the adults are ready to go trick or treating.	the kids are also ready to go get their loot.	they look adorable.	it was a busy night for those passing out candy.	what a great evening.
MAGIC	[male] came to a costume party he wasn't invited to.	many people showed up after that.	there were fireworks in the sky.	this was the highlight of the night.	my friend and i sat on the couch and talked for a long time.
Ours	it's halloween and a group of friends are dressed up for the party.	the girl is dressed as a witch and the guy is dressed as a zombie.	the party goes on and everyone is having a great time.	after the party, everyone goes to the haunted house.	at the end of the night, everyone drinks and has a great time.

GT	a group of family members have a great dinner.	the brother of the family is anxious for dinner to start.	he has a hearty soup with cream on it.	a plate of deserts are served at the end.	the family poses for a photo together.
MAGIC	[male] invited us to dinner a few rows down the street.	we started with steaks and onion rings.	here's the waiter standing at the table serving beer.	of course, dinner was served hot from the grill.	all in all, it was a great night.
Ours	i invited all of my friends over for dinner.	we had a great time talking and eating.	some of the food was very good.	at the end of the night, we all got together for a group photo.	it was a great night.

Figure 7: Qualitative comparison between our method and MAGIC. Words highlighted in red represents exact description of corresponding image, and words highlighted in green represents information from other images.

5.3 Qualitative Results

Fig. 7 presents two examples of generated stories by MAGIC and our proposed method. The results show that our approach generates stories with more accurate semantics that correspond to the images, as indicated by the red highlights. Moreover, the visual condition planner enables the generation of sentences that are relevant to the other images in the input sequence, as shown by the green highlights. Our method outperforms the baseline method in capturing visual contents within a single image and maintaining the theme of the album, resulting in stories of higher quality.

6 CONCLUSION AND DISCUSSION

In this paper, we propose a novel approach for visual storytelling that only requires textual story data for training. By leveraging the capabilities of pretrained cross-modality models such as CLIP, we model the visual storytelling task as a visual-conditioned generation problem. We adopt a guided decoding paradigm and design

a visual condition planner to aggregate the input visual sequence. Our method is evaluated on the VIST benchmark through extensive experiments, which demonstrate its effectiveness in generating high-quality visual stories.

Although the proposed method avert the cost of cross-modality annotated data, the training-free visual condition planner does have its limitations in understanding the complex temporal structures of visual input, which may affect the complexity of the generated story. In future work, it may be worth exploring few-shot learning methods to aggregate aligning results of image sequence to generate more informative and narrative stories.

ACKNOWLEDGMENTS

REFERENCES

[1] Ting-Hao 'Kenneth' Huang, Francis Ferraro, N. Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In *North American Chapter of the Association for Computational Linguistics*, 2016.

[2] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. Language models for image captioning: The quirks and what works. In *Annual Meeting of the Association for Computational Linguistics*, 2015.

[3] Pengcheng Yang, Fuli Luo, Peng Chen, Lei Li, Zhiyi Yin, Xiaodong He, and Xu Sun. Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling. In *International Joint Conference on Artificial Intelligence*, pages 5356–5362, 2019.

[4] Ruize Wang, Zhongyu Wei, Piji Li, Qi Zhang, and Xuanjing Huang. Storytelling from an image stream using scene graphs. *AAAI Conference on Artificial Intelligence*, pages 9185–9192, 2020.

[5] Youngjae Yu, Jiwan Chung, Heeseung Yun, Jongseok Kim, and Gunhee Kim. Transitional adaptation of pretrained models for visual storytelling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12653–12663, 2021.

[6] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

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

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901, 2020.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139, pages 8748–8763, 2021.

[10] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928, 2022.

[11] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected CLIP. In *Findings of the Association for Computational Linguistics*, pages 4055–4063, 2022.

[12] 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. In *AAAI Conference on Artificial Intelligence*, page 1445–1452, 2017.

[13] Diana Gonzalez-Rico and Gibran Fuentes Pineda. Contextualize, show and tell: A neural visual storyteller. *arXiv preprint*, abs/1806.00738, 2018.

[14] Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Annual Meeting of the Association for Computational Linguistics*, pages 899–909, 2018.

[15] Licheng Yu, Mohit Bansal, and Tamara Berg. Hierarchically-attentive RNN for album summarization and storytelling. In *Conference on Empirical Methods in Natural Language Processing*, pages 966–971, 2017.

[16] Qiuyuan Huang, Zhe Gan, Asli Celikyilmaz, Dapeng Oliver Wu, Jianfeng Wang, and Xiaodong He. Hierarchically structured reinforcement learning for topically coherent visual story generation. *AAAI Conference on Artificial Intelligence*, pages 8465–8472, 2019.

[17] Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao 'Kenneth' Huang, and Lun-Wei Ku. Knowledge-enriched visual storytelling. *AAAI Conference on Artificial Intelligence*, pages 7952–7960, 2020.

[18] Tengpeng Li, Hanli Wang, Bin He, and Chang Wen Chen. Knowledge-enriched attention network with group-wise semantic for visual storytelling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–12, 2022.

[19] Jiacheng Li, Haizhou Shi, Siliang Tang, Fei Wu, and Yueting Zhuang. Informative visual storytelling with cross-modal rules. In *ACM International Conference on Multimedia*, page 2314–2322, 2019.

[20] Tom Braude, Idan Schwartz, Alex Schwing, and Ariel Shamir. Ordered attention for coherent visual storytelling. In *ACM International Conference on Multimedia*, page 3310–3318, 2022.

[21] Mengshi Qi, Jie Qin, Di Huang, Zhiqiang Shen, Yi Yang, and Jiebo Luo. Latent memory-augmented graph transformer for visual storytelling. In *ACM International Conference on Multimedia*, page 4892–4901, 2021.

[22] Kohei Uehara, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. ViNTER: Image narrative generation with emotion-arc-aware transformer. In *Companion Proceedings of the Web Conference*, page 716–725, 2022.

[23] Ruichao Fan, Hanli Wang, Jinjing Gu, and Xianhui Liu. Visual storytelling with hierarchical bert semantic guidance. In *ACM Multimedia Asia*, 2022.

[24] Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. Exploring controllable text generation techniques. In *International Conference on Computational Linguistics*, pages 1–14, 2020.

[25] Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. DART: Open-domain structured data record to text generation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, 2021.

[26] Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych. Structural adapters in pretrained language models for AMR-to-Text generation. In *Conference on Empirical Methods in Natural Language Processing*, pages 4269–4282, 2021.

[27] Xu Zou, Da Yin, Qingyang Zhong, Hongxia Yang, Zhilin Yang, and Jie Tang. Controllable generation from pre-trained language models via inverse prompting. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 2450–2460, 2021.

[28] Fredrik Carlsson, Joey Öhman, Fangyu Liu, Severine Verlinden, Joakim Nivre, and Magnus Sahlgren. Fine-grained controllable text generation using non-residual prompting. In *Annual Meeting of the Association for Computational Linguistics*, pages 6837–6857, 2022.

[29] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint*, abs/1909.05858, 2019.

[30] Yizhe Zhang, Guoyin Wang, Chunyuan Li, Zhe Gan, Chris Brockett, and Bill Dolan. POINTER: Constrained progressive text generation via insertion-based generative pre-training. In *Conference on Empirical Methods in Natural Language Processing*, pages 8649–8670, 2020.

[31] Alvin Chan, Yew-Soon Ong, Bill Pung, Aston Zhang, and Jie Fu. Cocon: A self-supervised approach for controlled text generation. In *International Conference on Learning Representations*, 2021.

[32] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint*, abs/1912.02164, 2019.

[33] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics*, pages 4929–4952, 2021.

[34] Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Annual Meeting of the Association for Computational Linguistics*, pages 6691–6706, 2021.

[35] Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, 2021.

[36] Zhiyu Lin and Mark O. Riedl. Plug-and-blend: A framework for plug-and-play controllable story generation with sketches. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2021.

[37] Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics*, pages 3973–3997, 2021.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *International Conference on Neural Information Processing Systems*, page 6000–6010, 2017.

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, pages 1–67, 2020.

[40] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[41] Andrea Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. DeViSE: A deep visual-semantic embedding model. In *International Conference on Neural Information Processing Systems*, page 2121–2129, 2013.

[42] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision*, pages 67–84, 2016.

[43] Lluis Gomez, Yash Patel, Marçal Rusiñol, Dimosthenis Karatzas, and C. V. Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2017–2026, 2017.

[44] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. Learning visual n-grams from web data. In *IEEE International Conference on Computer Vision*, pages 4193–4202, 2017.

[45] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

[46] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint*, abs/2301.12597, 2023.

[47] Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. A contrastive framework for neural text generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 21548–21561, 2022.

[48] Yunjae Jung, Dahun Kim, Sanghyun Woo, Kyungsu Kim, Sungjin Kim, and In-So Kweon. Hide-and-tell: Learning to bridge photo streams for visual storytelling. In *AAAI Conference on Artificial Intelligence*, pages 11213–11220, 2020.

[49] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint*, abs/2205.02655, 2022.

[50] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005.

[51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[52] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

[53] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

[54] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, 2018.

[55] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. 2020.

[56] 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 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, 2016.

[57] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, 2016.