StoryLLaVA: Enhancing Visual Storytelling with Multi-Modal Large Language Models

Li Yang[1], **Zhiding Xiao**[1], **Wenxin Huang**[1,*], **Xian Zhong**[2]
[1] Hubei Key Laboratory of Big Data Intelligent Analysis and Application,
School of Computer Science and Information Engineering, Hubei University
[2] Hubei Key Laboratory of Transportation Internet of Things,
School of Computer Science and Artificial Intelligence, Wuhan University of Technology
yangli_lc@hubu.edu.cn, xiaozhiding_wh@stu.hubu.edu.cn,
wenxinhuang_wh@163.com, zhongx@whut.edu.cn

Abstract

The rapid development of multi-modal large language models (MLLMs) has positioned visual storytelling as a crucial area in content creation. However, existing models often struggle to maintain temporal, spatial, and narrative coherence across image sequences and frequently lack the depth and engagement of human-authored stories. To address these challenges, we propose Story with Large Language and Vision Assistant (StoryLLaVA), a novel framework for enhancing visual storytelling. Our approach introduces a Topic-Driven Narrative Optimizer (TDNO) that improves both the training data and MLLM models by integrating image descriptions, topic generation, and GPT-4-based refinements. Furthermore, we employ a preference-based ranked story sampling method that aligns model outputs with human storytelling preferences through positive-negative pairing. These two phases of the framework differ in their training methods: the former uses supervised fine-tuning, while the latter incorporates reinforcement learning with positive and negative sample pairs. Experimental results demonstrate that StoryLLaVA outperforms current models in visual relevance, coherence, and fluency, with LLM-based evaluations confirming the generation of richer and more engaging narratives. The enhanced dataset and model are available at https://github.com/XxxZzD/StoryLLaVA.

1 Introduction

Visual storytelling has seen significant progress with the rise of multi-modal large language models (MLLMs), particularly in narrative creation, screenplay writing, and content generation (Hao et al., 2022; Song et al., 2024; Yang et al., 2024; Zhang et al., 2024a). As modern content creation increasingly demands rich and thematically cohesive narratives, maintaining coherence and narrative appeal is crucial for effective visual storytelling. However, handling multiple images across diverse temporal, spatial, and plot dimensions presents challenges, particularly in ensuring consistency and engagement throughout the story.

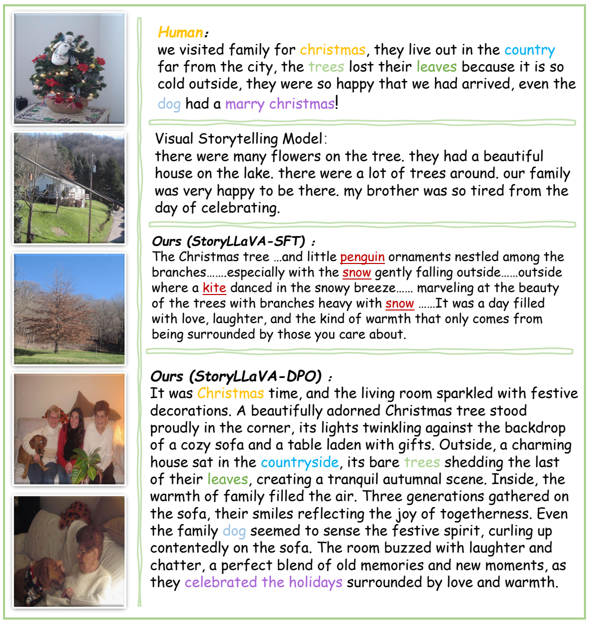


Figure 1: **Narrative examples generated by StoryLLaVA, compared to human-written stories and other visual storytelling models.** Words highlighted in the same color indicate semantic matches, while red words represent hallucinations.

Previous visual storytelling methods (Wang et al., 2018; Yu et al., 2021; Yang and Jin, 2023) typically employ neural networks to extract features from image sequences for end-to-end story generation. While these models perform well on automatic metrics, they struggle to match the engagement and diversity of human-authored stories. Some approaches (Chen et al., 2021; Hsu et al., 2021) enhance story diversity by combining object detection with narrative generation. Recent advancements in large MLLMs, such as Large Language and Vision Assistant (LLaVA) (Liu et al., 2023) and LLaVA-NeXT (Li et al., 2024a), extend storytelling capabilities to multi-image, video,

Proceedings of the 31st International Conference on Computational Linguistics, pages 3936–3951
January 19–24, 2025. ©2025 Association for Computational Linguistics

and 3D scenarios. Models like Generative Pre-trained Transformer (GPT)-4 (OpenAI, 2023) and Claude (Wu et al., 2023) approach human-level performance across various domains. However, challenges remain in generating high-quality, engaging stories from multi-image inputs, as current models often produce hallucinations that diverge from image content and fail to capture the richness of human-authored narratives, as illustrated in Figure 1.

High-quality data is crucial for generating coherent, expressive narratives with minimal hallucinations. However, directly using GPT-4 to augment story datasets often introduces irrelevant details, as it is not optimized for interpreting image sequences. Building on the success of image description models (Xiao et al., 2024), topic generation techniques (Pham et al., 2024), and GPT-4, we explore a Topic-Driven Narrative Optimizer (TDNO) within the LLaVA framework. This approach refines image descriptions, providing clearer guidance and context to improve dataset quality and benefit the Supervised Fine-Tuning (SFT) process. Through this method, we achieve more coherent, richer narratives that better align with sequential multi-image understanding.

Recognizing that pre-training and SFT alone are insufficient to eliminate hallucinations and improve generation quality, we propose a preference-based ranked story sampling method inspired by Direct Preference Optimization (DPO) (Zhang et al., 2024b). This method aligns model outputs with human storytelling preferences through positive-negative pairing of high- and low-quality stories. GPT-4 is used to score and rank these stories, constructing a preference dataset. This enables the model to learn narrative characteristics preferred by humans, enhancing coherence, consistency.

Building on the above motivations, we introduce Story with Large Language and Vision Assistant (StoryLLaVA), a novel framework structured around three phases: pre-trained StoryLLaVA, Storytelling with SFT (StoryLLaVA-SFT), and Storytelling with DPO (StoryLLaVA-DPO). Experimental results demonstrate that StoryLLaVA outperforms previous methods in visual relevance, coherence, and fluency. The model shows significant generalization improvements after multi-phase training, delivering high-quality results even on datasets not used during training.

The main contributions are fourfold:

- We introduce Story with Large Language and Vision Assistant (StoryLLaVA), a novel storytelling framework that generates more engaging and human-preferred narratives through a multi-phase learning strategy, enhancing both visual relevance and coherence.

- We propose a Topic-Driven Narrative Optimizer (TDNO) that enhances story datasets by combining image descriptions, topic generation, and GPT-4 refinements, addressing the shortage of high-quality training data.

- We implement a preference-based ranked story sampling strategy with positive-negative pairing to better align model outputs with human storytelling expectations.

- Our model consistently outperforms existing methods in both objective metrics and LLM-based evaluations, setting a new benchmark for visual storytelling.

2 Related Work

2.1 Visual Storytelling

Visual storytelling (Huang et al., 2016) focuses on generating coherent, human-like narratives from sequences of ordered images. This field introduced VIST, which has inspired significant advancements. Various frameworks have emerged (Kim, 2015; Yu et al., 2021), utilizing models based on convolutional neural networks (CNNs) combined with recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and Transformer-based architectures. Hsu et al. (2021); Xu et al. (2021) enhance story richness and diversity by incorporating external knowledge. Other approaches focus on specific aspects, such as sentiment (Chen et al., 2022), topics (Chen et al., 2024b), and textual style (Yang and Jin, 2023), improving coherence and richness across multiple dimensions.

Due to the scarcity of high-quality datasets, Ravi et al. (2021) introduced the AESOP dataset, featuring synthetic image sequences enriched with textual information. Similarly, Hong et al. (2023) presented VWP, composed of carefully curated movie frame sequences. However, the high cost of crowdsourcing for these datasets limits the scalability of high-quality data collection. To address this, we propose a Topic-Driven Narrative Optimizer (TDNO) that enriches datasets while ensuring thematic consistency and generating stable, high-quality story data.

2.2 Multi-modal Large Language Models

The development of large language models (LLMs) has accelerated research in multi-modal LLMs (MLLMs), which integrate multiple modalities. A commonly adopted architecture consists of: 1) a pre-trained visual encoder, 2) a pre-trained LLM, and 3) a multi-modal projector. For example, the Bootstrapping Language-Image Pre-training (BLIP) series, including BLIP-2 (Li et al., 2023a), uses a Q-Former to bridge the frozen LLM and vision encoder, while InstructBLIP incorporates task-specific visual guidance within the Q-Former. The LLaVA (Liu et al., 2023) series employs a simple linear layer as the multi-modal connector, followed by instruction fine-tuning. LLaVA-1.5 (Liu et al., 2024) improves this by segmenting input images, extracting key visual features, and integrating them with the original data for a more nuanced understanding. These models excel in tasks such as single-image description generation and Visual Question Answering (VQA). LLaVA-NeXT-Interleave (Li et al., 2024a) extends these capabilities to multi-image and multi-task transfer learning, though challenges remain in handling sequential multi-image tasks.

Regarding training strategies, preference alignment has become essential for improving LLM performance. Building on GPT-3, Brown et al. (2020) introduced InstructGPT, which incorporates reinforcement learning from human feedback (RLHF), outperforming GPT-3 despite using fewer parameters. In multi-modal LLMs, LLaVA-RLHF (Sun et al., 2024) enhances factual understanding through fact-based reinforcement learning, while LLaVA-Hound (Zhang et al., 2024b) employs Direct Preference Optimization (DPO) to improve performance in video QA tasks. Although methods such as Li et al. (2023b); Zhong et al. (2023); Chen et al. (2024a) improve caption generation, exploration in creative tasks like story generation remains limited.

3 Proposed Method

Given a sequence of consecutive images $I = \{I_n\}_{n=1}^{N}$, the task of story generation involves extracting event information and generating a coherent, engaging, and factually consistent story s.

3.1 StoryLLaVA Framework

As shown in Figure 2, we adopt the LLaVA (Liu et al., 2024) framework, incorporating several design modifications tailored to our task. We utilize the pre-trained SigLIP-so400m (Zhai et al., 2023) as the visual encoder, processing images at a resolution of 384×384, which provides higher resolution compared to Contrastive Language-Image Pre-Training (CLIP) (Radford et al., 2021). For a sequence of images $I = \{I_n\}_{n=1}^{N}$, where N is the number of images, the visual encoder generates feature maps $Z_n \in \mathbb{R}^{h \times w \times D}$ for each image:

$$\begin{aligned} Z &= \{Z_n\}_{n=1}^{N}, \\ Z_n &= g_{\psi}\left(I_n\right), \end{aligned} \tag{1}$$

where g_{ψ} is the visual encoder, h and w represent the spatial dimensions of the feature maps, and D is the feature dimension (number of channels). These visual features are projected into the language model's embedding space through a two-layer multi-layer perceptron (MLP) p_{θ}:

$$\begin{aligned} H &= \{H_n\}_{n=1}^{N}, \\ H_n &= p_{\theta}\left(Z_n\right), \end{aligned} \tag{2}$$

where $H_n \in \mathbb{R}^{(h \times w) \times K}$ represents the visual tokens for the n-th image, and K is the dimensionality of the language model's embedding space.

SigLIP generates 729 visual tokens per image ($h \times w = 27 \times 27 = 729$). To handle multiple images while respecting the language model's maximum token length of 4,096, we adopt Phi-3 Mini-128k (Abdin et al., 2024), which supports up to 128k tokens, enabling efficient training with limited computational resources.

3.2 Topic-Driven Narrative Optimizer

In visual storytelling, narrative text demands greater diversity than conventional image captions. Building on the Multi-level Description Generation method (Li et al., 2024b) and Sequence Data Enhancement (Zang et al., 2024), we propose the Topic-Driven Narrative Optimizer (TDNO) to enhance storytelling data.

3.2.1 Caption Generation

We use a story dataset $\{(I^{(i)}, y^{(i)})\}_{i=1}^{M}$, pairing image sequences $I^{(i)}$ with corresponding factual stories $y^{(i)}$. Captions are generated using Florence-2 (Xiao et al., 2024), a unified visual foundation model. Segmentation is performed by SAM (Kirillov et al., 2023), followed by descriptive text generation for local regions. Captions are refined using a phrase-level strategy (Dong et al., 2024), resulting in the final captions $C^{(i)} = \{C_n^{(i)}\}_{n=1}^{N_i}$.

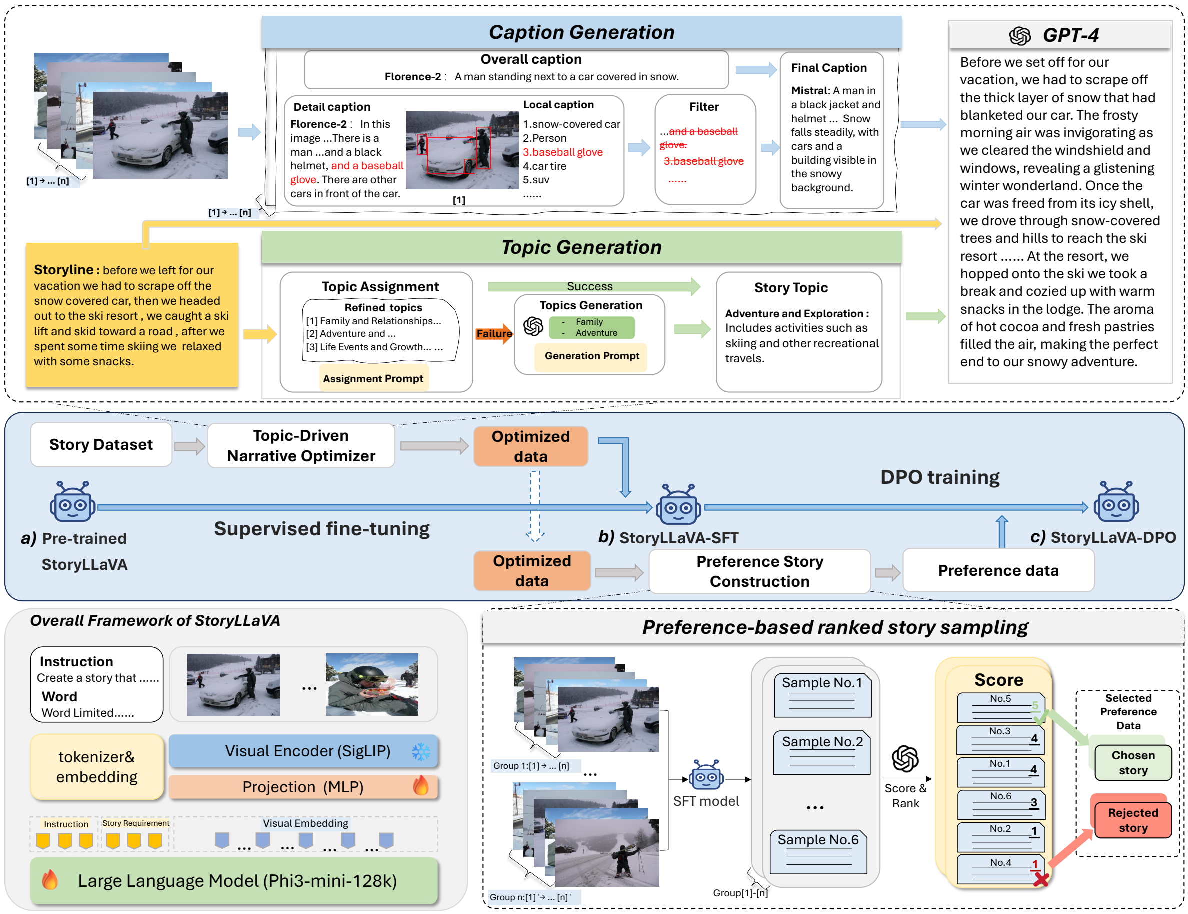


Figure 2: **Overview of the proposed pipeline.** Our framework contains three main phases: a) Phase 1: StoryLLaVA is pre-trained on LCS-558K. b) Phase 2: StoryLLaVA-SFT is fine-tuned with data from the TDNO. c) Phase 3: StoryLLaVA-DPO is trained using data from Preference Story Construction.

3.2.2 Topic Generation

We adapt TopicGPT (Pham et al., 2024) to extract and analyze story topics for our task. Example topics are manually written, and prompts are refined to guide GPT-4 (OpenAI, 2023) in generating relevant topics. After generating the initial topics, duplicates are merged, and rare topics are removed to ensure quality. To optimize computational efficiency, random sampling is employed to limit the number of processed entries.

Although GPT-4 generates high-quality topics, its cost limits scalability. To address this, we employ the open-source Mistral-7b (Jiang et al., 2023) model locally for topic assignment, significantly reducing API costs while maintaining accuracy. Refined topics are assigned to stories across datasets to ensure consistency with each story's content. For stories without assigned topics, new topics are generated based on manually selected examples. The topic generation process is formalized as:

$$T^{(i)} = \begin{cases} \text{Assign}\left(\mathcal{C}, y^{(i)}\right), & \text{if assignable,} \\ \text{Generate}\left(E, y^{(i)}\right), & \text{otherwise,} \end{cases} \tag{3}$$

where $T^{(i)}$ represents the topic for the i-th sequence, $\mathcal{C}$ is the generated topic pool, $y^{(i)}$ is the content of the i-th sequence, and E represents manually selected examples for generating new topics. Prompts for generation and assignment are provided in Appendix A.1.

3.2.3 Story Generation

Story generation integrates information from various sources while ensuring coherence and richness. We combine the original storyline $y^{(i)}$, image captions $C^{(i)}$, and the story topic $T^{(i)}$. Image captions provide scene-specific details, while the topic ensures thematic consistency, preventing deviations from the storyline.

The story generation process is formalized as:

$$S^{(i)} = \mathrm{GPT}\left(y^{(i)}, C^{(i)}, T^{(i)}, \mathrm{Prompt}\right), \quad (4)$$

where $y^{(i)}$ refers to the original storyline, $C^{(i)}$ are the image captions, and $T^{(i)}$ denotes the main topic. The Prompt guides GPT-4 (OpenAI, 2023) in merging these inputs into a coherent story $S^{(i)}$, ensuring alignment with the image sequence, length, and visual content, while avoiding off-topic deviations. The quantitative analysis results of the optimized story are shown in Appendix A.2.

3.3 Preference Story Construction

We identified three types of hallucinations in visual storytelling: 1) The narrative does not follow the actual image order, causing temporal incoherence. 2) The story mentions objects or elements not present in the images, leading to visual inconsistencies. 3) The story includes details not depicted in the images, causing deviations from the main storyline and topic.

We propose a story preference alignment method to mitigate hallucinations in MLLM-generated stories, addressing issues like incorrect visual information and narrative incoherence. Inspired by the Language Model Reward approach from LLaVA-Hound (Zhang et al., 2024b), we generate six stories from a set of ordered images, rearranging the image sequence to create additional samples. This process enhances the model's robustness to varied image sequences. During story generation, we set the temperature to 1.0, top-p to 0.2, and use a beam width of 4.

GPT-4 (OpenAI, 2023) evaluates the six sampled stories based on quality, reference story, and topic. The highest-scoring stories are selected as positive examples, while the lowest-scoring ones are chosen as negative examples. Both ranking and scoring ensure precise sample selection. Ranking identifies the best and worst-performing stories, while scoring ensures clear distinctions (samples are discarded if all scores are above or below 3). This contrast helps the model effectively distinguish between high- and low-quality outputs, improving its story generation ability. Scoring details are provided in Appendix A.3.

3.4 Training

3.4.1 Pre-training

Following the LLaVA (Liu et al., 2023) method, we pre-train the MLP connector by freezing the vision and language models and training on LCS-558K to align the two modalities effectively.

3.4.2 Supervised Fine-Tuning

We conduct Supervised Fine-Tuning (SFT) on an optimized dataset. Samples are selected from various datasets: 40k from VIST (Huang et al., 2016), 10k from VWP (Hong et al., 2023), 5k from PororoSV (Li et al., 2019), and 5k from FlintstonesSV (Maharana and Bansal, 2021), yielding 60k samples for fine-tuning. Additionally, we apply the LoRa (Hu et al., 2022) adapter to fine-tune the LLM.

3.4.3 DPO with Story Preference Data

Using the method described in Section 3.3, we construct 6,000 preference data entries, each consisting of an $\langle$ image sequence, chosen story, rejected story $\rangle$ sample for Direct Preference Optimization (DPO) training. The objective is defined as:

$$\begin{aligned} &\mathcal{L}_{\mathrm{DPO}}\left(\pi_\theta; \pi_{\mathrm{ref}}\right) \\ &= -\,\mathbb{E}_{(I,S_w,S_l)}\left[\log\sigma\left(\beta\begin{pmatrix}\log\frac{\pi_\theta(S_w|I)}{\pi_{\mathrm{ref}}(S_w|I)}, \\ -\log\frac{\pi_\theta(S_l|I)}{\pi_{\mathrm{ref}}(S_l|I)}\end{pmatrix}\right)\right], \end{aligned} \quad (5)$$

where S_w is the chosen story, S_l is the rejected story, I represents the image sequence, and $\mathbb{E}_{(I,S_w,S_l)}$ denotes the expectation over the dataset of image sequences and story pairs.

4 Experimental Results

4.1 Datasets

As outlined in Section 3, we utilize three datasets.

The first is LCS-558K image-caption dataset, curated by LLaVA (Liu et al., 2023), used for pre-training to align visual features.

The second dataset, employed in the Topic-Driven Narrative Optimizer (TDNO), combines four storytelling datasets: 1) Visual Storytelling (VIST) (Huang et al., 2016), containing sequences of five natural images from Flickr with corresponding five-sentence story descriptions; 2) Visual Writing Prompts (VWP) (Hong et al., 2023), consisting of film shots (5–10 images per sequence) with story; 3) PororoSV (Li et al., 2019), including one-second video clips, where keyframes serve as representative images, with multiple images forming a story unit; and 4) FlintstonesSV (Maharana and Bansal, 2021), featuring sequences of five consecutive video frames forming a story unit. Through the TDNO, we generate a total of 60k ⟨ image sequence, story ⟩ pairs.

The third dataset is the story preference dataset, with its preference data derived from the evaluation of the language model, containing 6k ⟨ image sequence, chosen story, rejected story ⟩ samples, constructed using the method described in Section 3.4.3.

4.2 Implementation Details

We employ the pre-trained SigLIP (Zhai et al., 2023) as the visual feature extractor in our framework. Due to resource constraints and the need to handle long-token inputs, we select the Phi-3-mini-128k (Abdin et al., 2024), a 3.8-billion-parameter model trained on 3.3 trillion tokens. Training proceeds in three stages: 1) Pre-training on LCS-558K for one epoch; 2) Supervised Fine-Tuning (SFT) on 60k samples over five epochs, with a learning rate of 2×10^{-4} and a batch size of 4; 3) Direct Preference Optimization (DPO) training on 6k preference data entries over three epochs, with a learning rate of 5×10^{-7} and a batch size of 8. Experiments are conducted on two NVIDIA A100 40GB GPUs and two NVIDIA A5000 24GB GPUs.

4.3 Baselines

We compare our model against the following baselines across various scenarios: 1) **Visual Storytelling Methods:** AREL (Wang et al., 2018), MCSM (Chen et al., 2021), and TAPM (Yu et al., 2021); 2) **Multi-modal Pipelines:** LLaVA (Liu et al., 2023) and BLIP2 (Li et al., 2023a), which generate image captions used as prompts for GPT-4 (OpenAI, 2023) to create stories; 3) **MLLMs:** LLaVA*, which modifies LLaVA's visual embedding module by concatenating embeddings from five images for fine-tuning. Details on instruction tuning are provided in Appendix A.6.

4.4 Evaluation Metrics

4.4.1 Automatic Evaluation

We adopt reference-free evaluation metrics (Yang et al., 2024; Surikuchi et al., 2024; Zeng et al., 2024) to assess the quality of generated stories. These metrics include GrooVIST (Surikuchi et al., 2023) for visual relevance, RoVIST-C (Wang et al., 2022) for story coherence, RoVIST-NR (Wang et al., 2022) for non-redundancy, and Intra-Repetition (Yao et al., 2019) for text fluency.

For VIST challenge, we use TAPM's (Yu et al., 2021) experimental setup and evaluate using reference-based metrics: CIDEr (C) (Vedantam et al., 2015), METEOR (M) (Banerjee and Lavie, 2005), and ROUGE-L (R) (Lin, 2004).

4.4.2 Human Evaluation

To further assess the quality of the stories generated by our system, we conducted a human evaluation. Each participant was provided with a questionnaire consisting of four sections. Each section contained 10 segments, with each segment comprising five consecutive images randomly sampled from the same dataset. Participants were then presented with three stories generated by 1) LLaVA*, 2) LLaVA + GPT-4, and 3) our StoryLLaVA-DPO model. Participants rated the generated stories on the dimensions of Relevance (Rel.), Attractiveness (Attr.), and Coherence (Coh.) using a 3-point Likert scale (Joshi et al., 2015). The detailed scoring method is provided in Appendix A.4.

4.4.3 LLMs Evaluation

Previous studies relied on human evaluation for image story assessment. However, human evaluations often lack stability due to subjective biases and emotional influences, particularly on a small scale (Clark et al., 2021). Recent research (Ning et al., 2023; Chhun et al., 2024) has explored using LLMs for visual storytelling evaluation, with findings indicating that GPT-4 (OpenAI, 2023)'s results align closely with human preferences (Bai et al., 2023; Liang et al., 2024).

For the LLM evaluation, we increased the sample size to 50 samples per dataset and employed GPT-4o and Claude 3.5 Sonnet. Notably, the "Relevance" dimension evaluates the relation between image captions and generated stories rather than the direct association between images and stories. Scoring details are provided in Appendix A.5.

4.5 Evaluation Results

4.5.1 Automatic Evaluation Results

Given the complexities of story generation, our goal is to produce stories that resemble human-authored narratives while reflecting diversity rather than adhering to a single Ground Truth. Traditional reference-based metrics often fail to capture the quality and coherence of generated content, particularly after the application of the TDNO.

Table 1 summarizes the results of reference-free automatic evaluation metrics applied to VIST test set, using five consecutive images without Ground Truth annotations. Our method consistently out-

	Methods	Metrics				Avg Story Len
		GROOVIST	RoVIST-C	RoVIST-NR	Intra-Repetition	
VST	AREL (Wang et al., 2018)	0.584	0.577	0.833	23.5	39.1
	MCSM + BART (Chen et al., 2021)	**0.852**	0.666	0.865	2.8	56.7
	TAPM (Yu et al., 2021)	0.734	0.671	<u>0.903</u>	6.8	45.0
Multi-Model	BLIP2 (Li et al., 2023a) + GPT-4	0.556	0.722	0.871	<u>1.2</u>	175.5
	LLaVA (Liu et al., 2023) + GPT-4	0.653	0.759	0.810	1.4	179.2
MLLMs	LLaVA* *w/* SFT	0.541	<u>0.809</u>	0.851	6.5	171.9
	a.StoryLLaVA (Pre-trained Only)	0.357	0.189	0.200	18.1	40.5
	b.StoryLLaVA *w/*SFT	0.578	0.772	0.856	3.4	170.7
	StoryLLaVA *w/* DPO (Ours)	<u>0.764</u>	**0.833**	**0.905**	**0.5**	160.6

Table 1: **Performance comparison of various visual storytelling models on VIST test set,** evaluated using reference-free metrics: GROOVIST (visual grounding), RoVIST-C (story coherence), RoVIST-NR (non-redundancy), and Intra-Repetition (sentence-level repetition). **Bold** and <u>underlined</u> values represent the best and second-best results, respectively.

Methods	M	R	C
AREL (Wang et al., 2018)	<u>35.2</u>	29.3	9.1
MCSM + RNN (Chen et al., 2021)	**36.1**	30.7	11.0
TAPM (Yu et al., 2021)	33.1	**37.2**	<u>13.8</u>
StoryLLaVA-SFT (Ours)	29.9	<u>33.7</u>	**14.5**

Table 2: **Performance on VIST test set using standard evaluation metrics. Bold** and <u>underlined</u> values represent the best and second-best results, respectively.

Method	Rel.	Attr.	Coh.	GM
LLaVA*	1.80	2.05	1.88	1.89
LLaVA + GPT4	2.25	**2.51**	2.20	2.31
StoryLLaVA *w/* DPO (Ours)	**2.37**	2.42	**2.29**	**2.38**

Table 3: **Human evaluation across three dimensions.** Geometric Mean (GM) represents overall performance, with all metrics rated on a scale from 0 to 3. **Bold** values represent the best results.

performs baselines in coherence, non-redundancy, and fluency. Notably, the MCSM + BART (Chen et al., 2021) model achieves the highest visual relevance score. The performance of multi-modal pipelines heavily depends on the output quality of caption models and prompt design. Improving MLLM comprehension capabilities could further enhance generation quality. Both our model and LLaVA (Liu et al., 2023) demonstrate the potential of MLLMs for multi-image story generation, though challenges remain in visual relevance and fluency. Ablation studies demonstrate that SFT is effective, and models further trained with DPO show significant improvements in visual relevance, fluency, and coherence.

Additionally, we conduct a reference-based evaluation for VIST challenge, with the Story with Large Language and Vision Assistant (StoryLLaVA) model fine-tuned exclusively on VIST under strict length constraints: "Ensure the story does not exceed five sentences". Table 2 presents these results, highlighting the effectiveness of MLLMs on this benchmark dataset.

4.5.2 Human Evaluation Results

Table 3 presents the results of the human evaluation. The performance of LLaVA* is suboptimal, primarily due to the generation of hallucinated information, which negatively impacts both visual relevance and story coherence. Leveraging the power of LLMs, LLaVA + GPT-4 achieves the highest attractiveness rating. However, its reliance on descriptions generated by LLaVA (Liu et al., 2023) often results in content misalignment, and the story generation prompt introduces instability in the output. In contrast, our method excels in both visual relevance and coherence, providing more consistent and accurate results.

4.5.3 LLMs Evaluation Results

Table 4 presents the results across four datasets, evaluated by GPT-4o (OpenAI, 2023) and Claude 3.5 Sonnet. Our method consistently outperforms LLaVA* and LLaVA + GPT-4 across most evaluation dimensions, excelling in relevance, coherence, and overall performance (GM). On VIST and VWP, our method demonstrates exceptional performance across all metrics, significantly surpassing other approaches, especially in maintaining relevance and logical coherence with the images. However, on PororoSV and FlintstonesSV datasets, our method shows relatively weaker performance. This may be attributed to the smaller

Dataset	Methods	Relevance		Attractiveness		Coherence		Geometric Mean	
		GPT-4o	Claude 3.5	GPT-4o	Claude 3.5	GPT-4o	Claude 3.5	GPT-4o	Claude 3.5
VIST	LLaVA*	2.39	2.33	2.65	**2.62**	2.35	2.47	2.46	2.47
	LLaVA + GPT-4	2.33	2.47	2.30	2.47	2.42	2.55	2.35	2.49
	StoryLLaVA *w/* DPO	**2.62**	**2.66**	**2.87**	2.61	**2.70**	**2.78**	**2.73**	**2.68**
VWP	LLaVA*	2.30	2.32	2.38	2.25	2.39	2.58	2.36	2.38
	LLaVA + GPT-4	2.27	2.36	2.45	2.40	2.52	2.57	2.41	2.44
	StoryLLaVA *w/* DPO	**2.49**	**2.50**	**2.59**	**2.45**	**2.88**	**2.68**	**2.65**	**2.54**
PororoSV	LLaVA*	1.88	2.12	**2.50**	2.39	2.52	2.57	2.25	2.35
	LLaVA + GPT-4	**2.40**	2.29	2.48	2.52	2.53	2.45	**2.47**	2.42
	StoryLLaVA *w/* DPO	2.18	**2.37**	2.46	**2.58**	**2.66**	**2.78**	2.44	**2.55**
FlintstonesSV	LLaVA*	1.97	2.10	2.45	2.50	2.47	2.39	2.26	2.32
	LLaVA + GPT-4	**2.38**	**2.44**	2.42	**2.62**	2.36	2.48	2.45	**2.51**
	StoryLLaVA *w/* DPO	2.23	2.32	**2.49**	2.58	**2.71**	**2.67**	**2.45**	2.50

Table 4: **Performance comparison of generated results across VIST, VWP, PororoSV, and FlintstonesSV evaluated by GPT-4o and Claude 3.5 Sonnet.** Geometric Mean represents overall performance, with all metrics rated on a scale from 0 to 3. **Bold** values represent the best results.

dataset size and lower image resolution (128×128), which could limit the model's ability to capture sufficient features during training, thus affecting generation quality. StoryLLaVA-DPO outperforms all other methods in both human evaluations and LLM-based assessments, indicating a consistent recognition of its high quality across the two evaluation approaches. In most metrics, LLaVA* scores lower than the other methods. Human evaluators tend to focus more on the practical utility and intuitive appeal of the results, whereas LLM evaluations are inclined to rely on patterns and linguistic features derived from training data.

4.6 Story Qualitative Analysis

Figure 3 presents qualitative results from three randomly selected test samples. On VIST (Huang et al., 2016) test images, compared to other models, our generated stories are imaginative and engaging. Human-created stories excel in emotional resonance and visual relevance, aligning closely with the image content. Although these stories often lack detailed descriptions, they effectively meet human expectations. In contrast, the stories generated by our model are longer, demonstrating stronger coherence, authenticity, and emotional depth. However, as the story length increases, the contextual relevance between the image and storyline gradually decreases.

Additionally, overly long texts may result in noticeable repetition. To address this, we limit the story length to 150–180 words, with the average storyline containing 30–36 words per image. For other models, shorter story lengths often lead to incomplete storylines, while those with similar lengths to ours tend to generate hallucinated information and perform poorly in visual relevance.

5 Discussions

Our approach enhances MLLM-based story generation by improving story length, coherence, and relevance, resulting in more engaging and vivid narratives. The Topic-Driven Narrative Optimizer (TDNO) enhances dataset quality by refining image descriptions and fostering more coherent topics. However, the dataset's diversity remains limited, which may affect the model's performance with dynamic or diverse content.

Future work could focus on diversifying the dataset or using active learning to enhance its quality. Although Direct Preference Optimization (DPO) has improved alignment with human preferences, integrating visual modality information into preference optimization could further strengthen the dataset and better reflect human judgment in visual storytelling. While our model can generate compelling narratives from image sequences, its ability to generalize to novel, unseen visual content remains a key challenge. Ensuring consistent and relevant narratives for any set of images, especially with varying visual styles, is an ongoing task.

6 Conclusion

In this work, we present Story with Large Language and Vision Assistant (StoryLLaVA), a novel framework for multi-modal story generation using MLLMs. By optimizing story data and employing a multi-phase training strategy, our model gen-

Figure 3: **Example stories generated by StoryLLaVA compared to baseline models, sampled from VIST, PororoSV, and FlintstonesSV.**

erates coherent, engaging, and human-preferred narratives from multi-image inputs. While our results are promising, challenges remain in reducing hallucinations and handling complex narratives. Future work will focus on improving robustness, ensuring visual consistency in longer stories, and enhancing preference alignment for more intricate plot structures. Additionally, we plan to explore the use of higher-resolution images and advanced multi-modal architectures to further improve story content and coherence.

6.1 Limitations

In optimizing the story data, we reduced costs by sampling datasets and using the open-source Mistral-7b (Jiang et al., 2023) model instead of GPT-3.5 for topic assignment. While this approach increased efficiency, it may have overlooked certain nuanced topics, especially when handling complex datasets, as Mistral-7b lacks the precision of GPT-3.5 in dealing with intricate cases.

For model selection under resource constraints, we opted for the Phi-3-mini-128k model due to its capability to handle multi-image inputs and long token sequences. While larger models, such as LLaMA3-8b (Meta, 2024), could provide better performance in processing image embeddings and generating more sophisticated stories, resource limitations influenced our choice.

Maintaining high-quality preference data remains a challenge in Direct Preference Optimization (DPO) training. Although we followed methods similar to LLaVA-Hound (Zhang et al., 2024b) to construct the preference dataset, this process still carries the risk of bias and struggles to fully reflect the diversity of human storytelling preferences.

Despite improvements, the model may still generate hallucinated details inconsistent with visual content, especially in low-resolution or incomplete scene descriptions. This highlights the need for stronger visual grounding mechanisms to address such issues.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62301213 and 62271361, the Hubei Provincial Key Research and Development Program under Grant 2024BAB039, and the Open Project Funding of the Hubei Key Laboratory of Big Data Intelligent Analysis and Application, Hubei University.

References

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv:2404.14219*.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023. TouchStone: Evaluating vision-language models by language models. *arXiv:2308.16890*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. Annu. Meet. Assoc. Comput. Linguistics Workshop*, pages 65–72.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Adv. Neural Inf. Process. Syst.*

Hong Chen, Yifei Huang, Hiroya Takamura, and Hideki Nakayama. 2021. Commonsense knowledge aware concept selection for diverse and informative visual storytelling. In *Proc. AAAI Conf. Artif. Intell.*, pages 999–1008.

Shuqin Chen, Xian Zhong, Yi Zhang, Lei Zhu, Ping Li, Xiaokang Yang, and Bin Sheng. 2024a. Action-aware linguistic skeleton optimization network for non-autoregressive video captioning. *ACM Trans. Multimedia Comput. Commun. Appl.*

Wei Chen, Xuefeng Liu, and Jianwei Niu. 2022. SentiStory: A multi-layered sentiment-aware generative model for visual storytelling. *IEEE Trans. Circuits Syst. Video Technol.*, 32(11):8051–8064.

Weiran Chen, Xin Li, Jiaqi Su, Guiqian Zhu, Ying Li, Yi Ji, and Chunping Liu. 2024b. TARN-VIST: Topic aware reinforcement network for visual storytelling. In *Proc. Joint Int. Conf. Comput. Linguistics*, pages 15617–15628.

Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. 2024. Do language models enjoy their own stories? prompting large language models for automatic story evaluation. *Trans. Assoc. Comput. Linguistics*, 12:1122–1142.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proc. Annu. Meet. Assoc. Comput. Linguistics*, pages 7282–7296.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. 2024. Benchmarking and improving detail image caption. *arXiv:2405.19092*.

Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. 2022. Language models are general-purpose interfaces. *arXiv:2206.06336*.

Xudong Hong, Asad B. Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual writing prompts: Character-grounded story generation with curated image sequences. *Trans. Assoc. Comput. Linguistics*, 11:565–581.

Chi-Yang Hsu, Yun-Wei Chu, Ting-Hao (Kenneth) Huang, and Lun-Wei Ku. 2021. Plot and rework: Modeling storylines for visual storytelling. In *Proc. Annu. Meet. Assoc. Comput. Linguistics Findings*, pages 4443–4453.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proc. Int. Conf. Learn. Representat.*

Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin 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. 2016. Visual storytelling. In *Proc. Conf. North Am. Chapter Assoc. Comput. Linguistics*, pages 1233–1239.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv:2310.06825*.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *Brit. J. Appl. Sci. Technol.*, 7(4):396–403.

Jeong-Hee Kim. 2015. *Understanding narrative inquiry: The crafting and analysis of stories as research*. Sage publications.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment anything. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 3992–4003.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. LLaVA-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv:2407.07895*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. Int. Conf. Mach. Learn.*, pages 19730–19742.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David E. Carlson, and Jianfeng Gao. 2019. StoryGAN: A sequential conditional GAN for story visualization. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 6329–6338.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024b. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 26763–26773.

Zipeng Li, Xian Zhong, Shuqin Chen, Wenxuan Liu, Wenxin Huang, and Lin Li. 2023b. Background disturbance mitigation for video captioning via entity-action relocation. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1–5.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. 2024. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Adv. Neural Inf. Process. Syst.*

Adyasha Maharana and Mohit Bansal. 2021. Integrating visuospatial, linguistic, and commonsense structure into story visualization. In *Proc. Conf. Empirical Methods Nat. Lang. Process.*, pages 6772–6786.

AI Meta. 2024. Introducing meta LLaMA 3: The most capable openly available llm to date. *Meta AI*.

Munan Ning, Yujia Xie, Dongdong Chen, Zeyin Song, Lu Yuan, Yonghong Tian, Qixiang Ye, and Li Yuan. 2023. Album storytelling with iterative story-aware captioning and large language models. *arXiv:2305.12943*.

OpenAI. 2023. GPT-4 technical report. *arXiv:2303.08774*.

Chau Pham, Alexander Miserlis Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proc. Conf. North Am. Chapter Assoc. Comput. Linguistics*, pages 2956–2984.

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. 2021. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Mach. Learn.*, pages 8748–8763.

Hareesh Ravi, Kushal Kafle, Scott Cohen, Jonathan Brandt, and Mubbasir Kapadia. 2021. AESOP: Abstract encoding of stories, objects, and pictures. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 2032–2043.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2024. MovieChat: From dense token to sparse memory for long video understanding. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 18221–18232.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *Proc. Annu. Meet. Assoc. Comput. Linguistics Findings*, pages 13088–13110.

Aditya K. Surikuchi, Sandro Pezzelle, and Raquel Fernández. 2023. GROOViST: A metric for grounding objects in visual storytelling. In *Proc. Conf. Empirical Methods Nat. Lang. Process.*, pages 3331–3339.

Aditya Kaushik Surikuchi, Raquel Fernández, and Sandro Pezzelle. 2024. Not (yet) the whole story: Evaluating visual storytelling requires more than measuring coherence, grounding, and repetition. In *Proc. Conf. Empirical Meth. Nat. Lang. Process. Findings*, pages 11597–11611.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 4566–4575.

Eileen Wang, Soyeon Caren Han, and Josiah Poon. 2022. RoViST: Learning robust metrics for visual storytelling. In *Proc. Conf. North Am. Chapter Assoc. Comput. Linguistics Findings*, pages 2691–2702.

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proc. Annu. Meet. Assoc. Comput. Linguistics*, pages 899–909.

Sean Wu, Michael Koo, Lesley Blum, Andy Black, Liyo Kao, Fabien Scalzo, and Ira Kurtz. 2023. A comparative study of open-source large language models, GPT-4 and claude 2: Multiple-choice test taking in nephrology. *arXiv:2308.04709*.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 4818–4829.

Chunpu Xu, Min Yang, Chengming Li, Ying Shen, Xiang Ao, and Ruifeng Xu. 2021. Imagine, reason and write: Visual storytelling with graph knowledge and relational reasoning. In *Proc. AAAI Conf. Artif. Intell.*, pages 3022–3029.

Dingyi Yang and Qin Jin. 2023. Attractive storyteller: Stylized visual storytelling with unpaired text. In *Proc. Annu. Meet. Assoc. Comput. Linguistics*, pages 11053–11066.

Dingyi Yang, Chunru Zhan, Ziheng Wang, Biao Wang, Tiezheng Ge, Bo Zheng, and Qin Jin. 2024. Synchronized video storytelling: Generating video narrations with structured storyline. In *Proc. Annu. Meet. Assoc. Comput. Linguistics*, pages 9479–9493.

Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proc. AAAI Conf. Artif. Intell.*, pages 7378–7385.

Youngjae Yu, Jiwan Chung, Heeseung Yun, Jongseok Kim, and Gunhee Kim. 2021. Transitional adaptation of pretrained models for visual storytelling. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 12658–12668.

Chuanqi Zang, Jiji Tang, Rongsheng Zhang, Zeng Zhao, Tangjie Lv, Mingtao Pei, and Wei Liang. 2024. Let storytelling tell vivid stories: An expressive and fluent multimodal storyteller. *arXiv:2403.07301*.

Jingjie Zeng, Liang Yang, Jiahao Kang, Yufeng Diao, Zhihao Yang, and Hongfei Lin. 2024. "barking up the right tree", a GAN-based pun generation model through semantic pruning. In *Proc. Joint Int. Conf. Comput. Linguistics*, pages 2119–2131.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 11941–11952.

Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2024a. MM-Narrator: Narrating long-form videos with multimodal in-context learning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 13647–13657.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, and Yiming Yang. 2024b. Direct preference optimization of video large multimodal models from language model reward. *arXiv:2404.01258*.

Xian Zhong, Zipeng Li, Shuqin Chen, Kui Jiang, Chen Chen, and Mang Ye. 2023. Refined semantic enhancement towards frequency diffusion for video captioning. In *Proc. AAAI Conf. Artif. Intell.*, pages 3724–3732.

A Appendix

A.1 Topic Generation and Assignment Prompts

We manually selected relevant example topics and refined prompts to guide GPT-4 (OpenAI, 2023) in generating appropriate topics. After generating the initial topics, duplicates were merged, and infrequent ones were removed to ensure quality and relevance. The Mistral-7b-Instruct-v0.3 (Jiang et al., 2023) model was used to assign the generated topics to the stories in the dataset. The final prompt displayed is the result of repeated refinement and testing for the story generation task, aiming to achieve a balance between complexity and specificity. The prompts for topic generation and assignment are shown in Figures 4 and 5.

A.2 Story optimization results

Table 5 displays the results of the evaluation before and after applying Topic-Driven Narrative Optimizer (TDNO) across the four datasets. We compared the differences in text features before and after story optimization across four datasets (VIST,

You will receive a document and a set of top-level topics from a topic hierarchy. Your task is to identify generalizable topics within the document that can act as top-level topics in the hierarchy. If any relevant topics are missing from the provided set, please add them. Otherwise, output the existing top-level topics as identified in the document.
[Top-level topics] {Topics}
[Examples]
Example 1: Adding "[1] Adventure and Exploration"
Document:
We found this tree when we were walking in a nearby town, it turns out it is a popular attraction here, the tree is very unusual, with its roots exposed, the trunk was really wide, as much as 12 feet! You can see how big these roots are—pretty amazing!
Your response:
[1] Adventure and Exploration: Involves exploring new places and discovering remarkable natural attractions.
Example 2: Duplicate "[1] Life Events and Growth", returning the existing topic
Document:
I took a nice stroll around the neighborhood, and this blossom of flowers let me know that spring is here officially. I felt like a little child looking through the lens of this binocular. This is the name of the park where I was taking my stroll. When I was growing up, I thought this statue was built wrong because it is missing a full head. The grass looks so well-manicured and taken care of.
Your response:
[1] Life Events and Growth: Involves personal reflections on experiences from childhood to adulthood and changes in perception over time.
[Instructions]
Step 1: Determine topics mentioned in the document.
- The topic labels must be as GENERALIZABLE as possible. They must not be document-specific.
- The topics must reflect a SINGLE topic instead of a combination of topics.
- The new topics must have a level number, a short general label, and a topic description.
- The topics must be broad enough to accommodate future subtopics.
Step 2: Perform ONE of the following operations:
1.If there are already duplicates or relevant topics in the hierarchy, output those topics and stop here.
2.If the document contains no topic, return "None".
3.Otherwise, add your topic as a top-level topic. Stop here and output the added topic(s). DO NOT add any additional levels.
[Document]
{Document}
Please ONLY return the relevant or modified topics at the top level in the hierarchy.
[Your response]

Figure 4: Prompt for topic generation.

VWP, PororoSV, and FlintstonesSV), focusing on three key metrics: average story length, type-token ratio (TTR), and the number of unique terms.

Method	Avg Story Len	TTR	Terms
VIST	51.90	0.76	39.24
VIST *w/* TDNO	157.68	0.70	112.77
VWP	68.60	0.72	53.77
VWP *w/* TDNO	188.20	0.72	136.98
PororoSV	65.52	0.63	38.49
PororoSV *w/* TDNO	164.17	0.72	118.75
FlintstonesSV	81.21	0.57	46.02
FlintstonesSV *w/* TDNO	165.30	0.71	117.26

Table 5: **Comparison of text features before and after TDNO optimization across four datasets.**

A.3 Building the Preference Dataset

Figure 6 provides a detailed explanation of the scoring process used to select positive and negative examples for Direct Preference Optimization (DPO) training. To enhance robustness, we rearranged image sequences to create alternative samples, exposing the model to varied input arrangements. When rearranging image sequences, the storyline was modified, and results were regenerated using the method described in Section 3.2 to establish new references. Surikuchi et al. (2024) highlighted three key factors for evaluating story quality: visual grounding, coherence, and repetitiveness. These criteria formed the basis for feedback scoring.

A.4 Details of Human Evaluation

All reviewers were university students from diverse fields and had no direct affiliation with the project, ensuring objectivity and fairness. A total of 15 members scored 40 image-story pairs. Reviewers, fluent in English, assessed the generated English stories for accuracy and clarity. Each reviewer conducted the assessment independently, ensuring fair and unbiased scoring. Following Yang et al. (2024), evaluations were based on three aspects: 1) **Relevance** (Rel): the association between generated stories and source images; 2) **Attractiveness** (Attr): the ability of stories to engage readers; 3) **Coherence** (Coh): the logical flow between story sentences. The geometric mean (GM) was used to measure overall performance, with scores ranging from 0 to 3 in each dimension.

A.5 Details of LLM Evaluation

We employed advanced closed-source large models, GPT-4o (OpenAI, 2023) and Claude 3.5 Sonnet, and evaluated them using the same scoring criteria as human assessments. Detailed scoring criteria are shown in Figure 7. An example of GPT-4o's evaluation is presented in Figure 8, which includes GPT-4o's analysis and scoring of relevance, attractiveness, and coherence.

You will receive a document and a topic hierarchy. Assign the document to the most relevant topics in the hierarchy. Then, output the topic labels, assignment reasoning, and supporting quotes from the document. DO NOT make up new topics or quotes.
[Topic Hierarchy]
{tree}
[Examples]
Example 1: Assign "[1] Adventure and Exploration" to the document
Document:
We found this tree when we were walking in a nearby town, it turns out it is a popular attraction here, the tree is very unusual, with its roots exposed, the trunk was really wide, as much as 12 feet! You can see how big these roots are—pretty amazing!
Assignment:
[1] Adventure and Exploration: Involves discovering a notable natural attraction ("...it is a popular attraction here...the tree is very unusual, with its roots exposed...")
Example 2: Assign "[1] Life Events and Growth" to the document
Document:
I took a nice stroll around the neighborhood, and this blossom of flowers let me know that spring is here officially. I felt like a little child looking through the lens of this binocular. This is the name of the park where I was taking my stroll. When I was growing up, I thought this statue was built wrong because it is missing a full head. The grass looks so well-manicured and taken care of.
Assignment:
[1] Life Events and Growth: Involves personal reflections on experiences and changes over time ("...When I was growing up, I thought this statue was built wrong...")
[Instructions]
1.Topic labels must be present in the provided topic hierarchy. You MUST NOT make up new topics.
2.The quote must be taken from the document. You MUST NOT make up quotes.
[Document]
{Document}
Double check that your assignment exists in the hierarchy!
Your response:

Figure 5: Prompt for topic assignment.

A.6 LLaVA Fine-tuning

We used version 1.5 of the LLaVA (Liu et al., 2023) model, incorporating CLIP-ViT-L-336px and LLaMA-7b. The model's visual embedding component was modified, and fine-tuning was performed using LoRA.

A.7 Additional Generated Examples

Figure 9 presents additional examples of stories generated by Story with Large Language and Vision Assistant (StoryLLaVA).

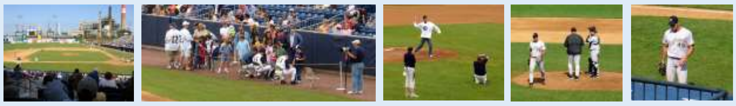
Given the following inputs:
1.Reference Story: {reference story}
2.Reference Story Topic: {Topic}
3.Six Model Predicted Stories: {prediction1}, {prediction2}, {prediction3}, {prediction4}, {prediction5}, {prediction6}
Your task is to evaluate each model-predicted story based on the following criteria:
-Visual Grounding: Does the predicted story accurately reflect visual descriptions or imagery consistent with the topic or prompt? Are the visual elements realistic and effectively grounded in the context of the story?
-Coherence: Does the predicted story flow logically from one scene to another? Evaluate whether the narrative is consistent and makes sense overall, with clear transitions and a structured storyline.
-Repetitiveness: Does the predicted story avoid unnecessary repetition of phrases, descriptions, or ideas? Check whether any redundant elements affect the smoothness or readability of the narrative.
For each story, output the following:
-Score: A quality score from 1-5 based on how well the story satisfies the criteria (higher is better).
-Rank: Rank the six predicted stories from best (1) to worst (6) based on the overall quality across the criteria.
Output Format:
For each predicted story:
Story {prediction number}:
-Score: <integer from 1-5>
-Rank: <rank from 1-6>

Figure 6: GPT evaluation prompt for story generation. The prompt includes a reference story, story topic, and model predictions, followed by a quality evaluation with corresponding scores and ranks based on predefined criteria.

Given the following inputs:
1.Image captions: {image captions}
2.Model Predicted Stories: {story1, story2, story3…}
Your task is to evaluate each model-predicted story:
Relevance: The degree of association between the generated stories and the source images.
0-1: The story has some connection to the source images but mostly overlooks key visual elements, resulting in poor relevance.
1-2: The story is mostly relevant to the source images, with only small parts being slightly off-topic or missing minor visual details.
2-3: The story is highly relevant to the source images, fully addressing and reflecting the visual content in detail.
Attractiveness: The ability of the stories to capture and maintain human interest.
0-1: The content has low attractiveness, generating some interest but struggling to maintain the reader's attention.
1-2: The content is quite engaging, able to generate interest and sustain the reader's attention for a substantial duration.
2-3: The content is highly engaging, immediately grabbing the reader's interest and maintaining their full attention through out.
Coherence: The extent to which the sentences in a story are logically connected.
0-1: The content has poor coherence, with noticeable jumps or breaks in logic, making the flow unclear.
1-2: The content is mostly coherent, with clear logical relationships between sentences and paragraphs, though there may be minor discontinuities.
2-3: The content is highly coherent, with a tightly organized logical structure that smoothly connects sentences and paragraphs in a clear and readable manner.
You must compare the stories, please apply strict standards to provide detailed scoring for each dimension, highlighting their strengths and weaknesses to better distinguish between them.
For each story, output the following:
Output Format:
Relevance -Score: <decimal from 0-3>
Attractiveness - Score: <decimal from 0-3>
Coherence - Score: <decimal from 0-3>

Figure 7: GPT-4o and Claude 3.5 evaluation prompt for story generation.

Images:



StoryLLaVA Story:
The stadium buzzed with excitement as fans poured into the stands, filling every seat. The air was electric with anticipation. On the field, players and coaches mingled with the crowd, signing autographs and posing for photos. The game was about to begin, and the energy was palpable. As the first pitch soared, the crowd erupted in cheers. Cameras flashed, capturing the moment as players and spectators alike shared in the thrill of the game. In the dugout, coaches strategized and gave last-minute pep talks, their voices blending with the roar of the crowd. The tension was high, every moment charged with the possibility of a game-changing play. Fans leaned forward in their seats, eyes wide with focus, as the players battled it out under the bright stadium lights. The game was more than just a contest; it was a shared experience, a tapestry of cheers, laughter, and heart-stopping moments woven together by the collective spirit of everyone present. As the innings passed, the excitement never waned, proving that baseball was indeed America's pastime, cherished by all who witnessed its magic firsthand.

Relevance:
Score:2.5
The story closely aligns with the images: it mentions the atmosphere in the stadium, the crowd, the players interacting with fans (seen in image 2), the game being played (images 5), and the high energy around the game. However, it could have included more specific details such as the pitcher and coaches (images 1, 3, and 4), making it slightly off-topic in certain moments.

Attractiveness:
Score:2.7
The story is engaging and captures the essence of a live baseball game experience, with vivid descriptions of crowd excitement, player interactions, and game tension. It successfully maintains the reader's interest from start to finish, highlighting the shared emotions and the thrill of the game.

Coherence:
Score: 2.5
The story is well-structured, flowing naturally from the pre-game excitement, through the pitch, to the intense gameplay. Each sentence builds on the previous one, creating a coherent and smooth narrative that makes sense from beginning to end.

Human Story:
this is a stadium that we saw our favorite team play baseball, they let us come on the field to meet some of the players, they even let me throw a pitch from the pitcher 's mound, we saw how the baseball officials worked, this was a great player that talked to us before we went home.

Relevance:
Score: 2.8
The story directly relates to the images, mentioning the experience of meeting players (image 2), throwing a pitch from the mound (image 1), and seeing baseball officials at work (images 3 and 4). The reference to interacting with players and throwing a pitch matches the visual content perfectly.

Attractiveness:
Score: 1.4
While the story is personal and touches on meaningful interactions, it lacks vivid descriptions and emotional depth that could make it more engaging. It states the facts but doesn't capture the excitement or detail of the day.

Coherence:
Score: 2.0
The story is generally coherent, but it feels a bit rushed and disjointed at points, such as jumping from throwing a pitch to meeting players and going home without much transition. A smoother flow between these events would improve overall coherence.

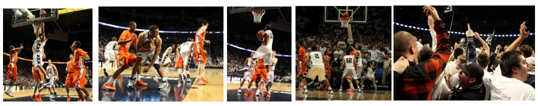
LLaVA+GPT4 Story:
On a hot summer day, excitement fills a bustling baseball field. A young pitcher, under the watchful eyes of a coach and a photographer, delivers a powerful throw, capturing the attention of everyone present. Nearby, players from the Long Island Ducks are signing autographs, spreading joy among a group of children and adults alike.
The game hits a critical moment as the pitcher discusses strategies with the catcher and coach on the mound, the tension palpable. Elsewhere, a player wearing jersey number 49 readies himself, his focus reflecting the intense atmosphere of the competition.
The climax of the game arrives as the pitcher executes a precise throw and the batter swings, with the crowd holding their breath in anticipation. This day's game is more than a sporting event; it's a display of community unity and the indomitable spirit of sports.

Relevance:
Score: 2.3
The story directly relates to the images, mentioning the experience of meeting players (image 2), throwing a pitch from the mound (image 1), and seeing baseball officials at work (images 3 and 4). The reference to interacting with players and throwing a pitch matches the visual content perfectly.

Attractiveness:
Score: 1.9
The story is engaging, successfully capturing the excitement of the baseball game. It keeps the reader interested and involved in the narrative, although there might be room to intensify the emotional impact or detail to reach a broader audience.

Coherence:
Score:2.3
The narrative is coherent and logically structured. The transitions between different scenes (pitching, strategizing, community involvement) are smooth, though there might be slight room for improvement in connecting these elements more tightly to enhance the flow.

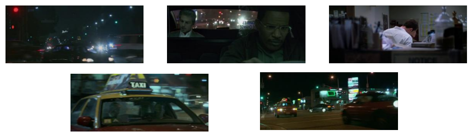
Figure 8: Example of an evaluation conducted by GPT-4o.



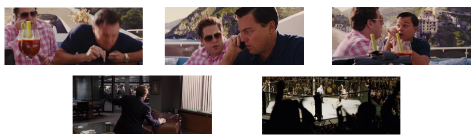
StoryLLaVA-DPO:
The basketball game was electrifying from the start. The orange team took an early lead, and the crowd roared as one of their players expertly dribbled past defenders. The white team fought back fiercely, gradually adjusting their rhythm and launching a powerful offensive. As the game progressed, a white team player broke through the defense and delivered a spectacular dunk, igniting the atmosphere in the arena. Despite their efforts, the orange team struggled to contain the relentless attacks from the white team. With precise coordination and strong offensive momentum, the white team eventually overtook the score. As the final buzzer sounded, the crowd erupted in cheers, celebrating the white team's victory. Although disappointed, the orange team showed respect, acknowledging the skill and determination of their opponents. It was a game to remember, filled with thrilling moments and excellent sportsmanship from both sides.

StoryLLaVA-DPO:
Today was the big day—her wedding! The venue was picturesque, with a grand arch adorned with flowers and a serene water body in the background. Her bridesmaids, all dressed in stunning red dresses, stood gracefully around her, their smiles reflecting the joy of the occasion. As she walked down the aisle, the crowd erupted in applause, and her father beamed with pride. The ceremony was intimate and beautiful, set against the backdrop of lush greenery and a clear blue sky. After exchanging vows, the newlyweds shared a heartfelt kiss, sealing their promises with love. The reception that followed was nothing short of spectacular. Guests mingled under a tent decorated with flowers, and laughter filled the air. The bride and groom posed for numerous photos, capturing every precious moment of their special day. By the end of the night, everyone agreed—it was a magical celebration filled with love, joy, and unforgettable memories.

(a) VIST



StoryLLaVA-DPO:
As the night descended, the city streets came alive with the glow of streetlights and the hum of engines. Cars whizzed by, their headlights cutting through the darkness. Inside one car, a man in a black shirt and sunglasses sat with a serious expression, his eyes fixed on the road ahead. The night was his alone, and he relished the quiet solitude of the drive. His thoughts wandered as he navigated the busy streets, filled with buildings, trees, and the occasional taxi. The city was a symphony of lights and motion, each car adding to the vibrant tapestry of the night. He passed by a yellow taxi, its lights blinking rhythmically, a reminder of the countless stories unfolding around him. As he drove deeper into the heart of the city, the buildings grew taller, their illuminated windows reflecting the dreams and lives within. The streets were a maze of possibilities, each turn promising new adventures. The night was endless, and so were the roads that stretched before him, inviting him to explore the boundless journey ahead.



StoryLLaVA-DPO:
Matt and Leonardo sat at a cozy restaurant table, their laughter filling the air as they reminisced about old times. The table was cluttered with remnants of their meal, but their focus was entirely on each other. Suddenly, Leonardo's phone buzzed. He glanced at the screen, his face turning serious. "I have to take this outside," he said, rising from his chair. Outside, the bustling city street provided a stark contrast to their intimate moment. Leonardo walked briskly to his car, his mind racing with the urgency of the call. As he drove, he couldn't shake the feeling that something significant was about to happen. The city lights blurred past him, but his thoughts remained fixed on the conversation he was about to have. The anticipation was palpable, and as he approached his destination, he braced himself for whatever news awaited him. Little did he know, this moment would mark the beginning of a new chapter in his life, one filled with unexpected twists and turns.

(b) VWP

Figure 9: Additional examples generated by StoryLLaVA.