🪄MAGIC: Multimodal Story Generation from Image Collections

Anonymous ACL submission

Abstract

We introduce a new task, **M**ultimod**a**l Story **G**eneration from **I**mage **C**ollections (MAGIC), where the goal is to generate a coherent narrative conditioned on a small subset of images that must be retrieved from a large, unordered collection. This underexplored task reflects real-world constraints where stories must be created using limited visual assets. To address the dual challenges of strategically selecting narratively useful images and constructing a coherent, visually grounded story, we propose a narrative-centric framework that first selects a diverse yet compatible subset of images, next infers their temporal and causal ordering, then bridges narrative gaps, and finally expands into a full story. We also build a dataset with 4,000 images to support this new task. Extensive automated and human evaluations show that our approach significantly outperforms baseline methods in narrative coherence, logical sense, novelty, naturalness, and visual engagement, establishing a strong foundation for multimodal storytelling under realistic resource constraints.

1 Introduction

Story generation has long been an important problem in natural language generation, aiming to automatically produce coherent, engaging, and contextually appropriate narratives from textual prompts (Clark et al., 2018; Fan et al., 2018, 2019; Yao et al., 2019; Yang et al., 2022, 2023b; Tian et al., 2024b). With the rapid progress of vision-language models (VLMs) (Ramesh et al., 2021; Alayrac et al., 2022; Li et al., 2022; Liu et al., 2023a; Achiam et al., 2023), recent years have witnessed a growing interest in multimodal storytelling where textual and visual modalities are jointly leveraged to create richer and more immersive storytelling experiences. Such systems have significant applications in education, entertainment, marketing, and digital content creation, enabling automated production of storybooks, interactive narratives, and visualized story summaries from multimodal inputs.

Despite substantial progress in multimodal storytelling, prior research has primarily focused on three dominant paradigms—visual storytelling (Huang et al., 2016), story visualization (Li et al., 2019), and interleaved image–text generation for storytelling (Tian et al., 2024a; Yang et al., 2024; An et al., 2024)—each of which operates under restrictive assumptions. Visual storytelling assumes that the input is a fixed, temporally ordered image sequence (e.g., consecutive photo albums or video frames) and focuses on generating a narrative to describe that sequence. Story visualization takes the opposite direction, presupposing that a complete textual storyline is already available and producing corresponding visual content for each event. Interleaved generation jointly produces images and text within a single pipeline, requiring simultaneous control over both modalities throughout the narrative. All these paradigms inherently rely on pre-defined inputs from one modality or unrestricted generative capabilities in both modalities, limiting their applicability in real-world scenarios where such assumptions might not always hold.

In this work, we propose a new and practically motivated task: **M**ultimod**a**l Story **G**eneration from **I**mage **C**ollections (MAGIC). Given a user query specifying a story intent, the system must select a small subset of images from a large, unordered image collection and compose a textual narrative that is grounded in those images. The resulting story should not only faithfully reflect the visual content of the selected images but also form a coherent, logically structured narrative that fulfills the user intent. This task formulation is inspired by real-world constraints frequently encountered in content creation workflows. Practical domains such as journalism, museums, education, and entertainment typically maintain proprietary image databases due to considerations of copyright, trade-

mark, or privacy (Crews and Brown, 2011; Taurino et al., 2025). Consequently, professional storytellers in these fields are often required to construct narratives exclusively from in-house visual assets rather than sourcing external new ones. In addition, computational or operational constraints may further limit the number of images that can be retrieved and processed, necessitating story generation from a small subset of available visuals.

The primary challenges in this task arise from the dual complexity of image selection and narrative construction. First, the system must determine which images to retrieve from the collection. Naïvely choosing the most relevant images often leads to visually repetitive scenes and redundant events, resulting in stories that are monotonous and less engaging. Second, the system must decide how to construct a compelling story from the selected images. Even when the images are individually meaningful, they may not naturally form a plausible narrative sequence. Bridging the gaps between visually depicted events and weaving them into a coherent, logically grounded storyline requires deeper reasoning about temporal order, causal dependencies, and narrative structure.

To address these challenges, we introduce a two-stage framework: Narrative-Centric Image Selection and Content Planning. In the first stage, we aim to select a small subset of images that are both diverse—covering a broad range of distinct narrative events—and compatible, ensuring they can be plausibly connected into a storyline. In the second stage, we infer a plausible narrative structure by analyzing temporal and causal relations among the selected events, arranging them into a coherent order. Finally, we bridge narrative gaps between events and expand into a full story, enriching transitions and ensuring story continuity.

We further emphasize that this is the first systematic study of multimodal story generation under a resource-constrained, collection-based setting. No existing work has investigated this problem, and no publicly available datasets support it. To facilitate research in this direction, we construct a new dataset comprising 40 topics, 200 user queries, and 4,000 images, specifically designed for daily-life, educational, and child-oriented storytelling scenarios—domains that represent some of the most common and impactful applications of story-based content generation.

We conduct extensive experiments to evaluate our approach, including automated evaluation with GPT-5 as a judge, human evaluation with pairwise comparisons, ablation studies to assess the contribution of each component, and fine-grained analyses to verify the effectiveness of core individual operations. Results show that our method substantially outperforms baseline methods across multiple evaluation dimensions, particularly in narrative coherence, logical sense, novelty, naturalness, and visual engagement.

In summary, our contributions are as follows:

- We propose a new task, *multimodal story generation from image collections*, which reflects practical constraints in real-world content creation and has not been explored in prior work.
- We introduce a narrative-centric framework that jointly addresses the challenges of image selection and story construction by maximizing event diversity, ensuring narrative compatibility, and inferring coherent story structures.
- We build a dataset tailored to this new task, enabling systematic study and evaluation.
- We demonstrate through comprehensive experiments that our method significantly improves story quality over baseline methods across both automated and human evaluations.

2 Related Work

2.1 Multimodal Storytelling

By jointly leveraging textual and visual modalities, multimodal storytelling systems aim to generate richer, more engaging narratives that go beyond purely textual story generation. Existing work in this area can be broadly categorized into three main paradigms: visual storytelling, story visualization, and interleaved image-text generation for storytelling. Visual storytelling (Huang et al., 2016; Yu et al., 2017; Wang et al., 2018a,b; Huang et al., 2019; Hsu et al., 2019; Hu et al., 2020; Wang et al., 2020; Chen et al., 2021; Xu et al., 2021; Hsu et al., 2021; Hong et al., 2023) focuses on generating textual narratives conditioned on a fixed sequence of images. These images are ordered—such as consecutive photographs or video frames—and inherently encode an implicit storyline. Story visualization (Li et al., 2019; Maharana et al., 2021; Maharana and Bansal, 2021; Maharana et al., 2022; Chen et al., 2022a; Gong et al., 2023; Rahman et al., 2023; Pan et al., 2024; Liu et al., 2024a; Shen et al., 2025; Shen and Elhoseiny, 2025) takes a textual narrative as input and generates a sequence of images, one for each sentence or event. In this paradigm,

the textual input already defines the storyline explicitly. More recently, interleaved image-text generation for storytelling (Tian et al., 2024a; Yang et al., 2024; An et al., 2024; Chen et al., 2025a; Zhou et al., 2025; Chen et al., 2025b; Kou et al., 2025) has gained attention, where both images and text are generated jointly within a single pipeline. This paradigm allows the model to dynamically co-adapt the two modalities throughout the narrative, producing multimodal stories in which visual and textual elements evolve together in a tightly coupled manner.

While these paradigms have significantly advanced multimodal storytelling, they all have underlying assumptions: visual storytelling presumes that the input images form a natural chronological sequence, story visualization relies on an existing storyline provided by text, and interleaved generation assumes full control over both modalities during generation. However, these assumptions may break down in some of the practical storytelling workflows, where available images are unordered and must be selected from a large proprietary database due to copyright, trademark, or privacy restrictions. In contrast, in our proposed task the system must first strategically retrieve a small, narratively useful subset of images from a large, unordered collection and then construct a coherent story based on those images. This new task more closely reflects the constraints faced in practical domains and also introduces new challenges, i.e., how to jointly optimize image selection and narrative construction under resource limitations.

2.2 Image Retrieval and Re-ranking

A related line of research focuses on retrieving images from large-scale collections for downstream tasks such as captioning (Liu et al., 2018), storytelling (Chen et al., 2019; Yang et al., 2023a), and question answering (Chen et al., 2022b; Feng et al., 2025). Most existing VLM-based approaches for image retrieval (Liu et al., 2021; Koh et al., 2023; Lin et al., 2025) rely on semantic similarity to rank and select images, optimizing for relevance to the input query. While effective for most retrieval tasks, this approach overlooks the interactions among retrieved images—a critical factor to consider when the goal is to construct a cohesive narrative. Selecting multiple visually similar or redundant images, for instance, can result in repetitive events and less engaging stories.

There have been attempts at re-ranking the retrieved images while taking into account the needs of different downstream tasks. For example, Tan et al. (2021) propose Re-ranking Transformers which incorporate both local and global visual features to re-rank the retrieved images for the purpose of geometric verification. Zhu et al. (2023) propose to re-rank retrieved images by reasoning over correlations between local patch pairs, attention values, and position/coordinate data, so as to decide if two images come from the same place for place recognition. In contrast, our method introduces a narrative-centric image selection perspective that jointly considers semantic relevance, event diversity, and narrative compatibility. By selecting images that are distinct yet compatible within a storyline, our approach ensures that the retrieved visual evidence forms a meaningful foundation for subsequent multimodal story generation.

3 Task Definition and Dataset Creation

3.1 Task Definition

We define the task of **multimodal story generation from image collections (MAGIC)** as follows: *given a user query requesting a story, the system aims to compose a textual narrative associated with a small subset of images that must be retrieved from a large image collection*. The generated story should not only align with the visual semantics but also form a continuous narrative that satisfies the user intent.

We focus on this formulation to closely mirror real-world storytelling scenarios where content creators usually operate under practical constraints and cannot assume unrestricted access to visual data. In many professional contexts—such as children's storybook production, educational content design, museum exhibit curation, and journalistic storytelling—narratives must be built around a limited set of in-house images. Computational or operational resources may further constrain the number of images that can be retrieved and processed, requiring stories to be generated from a small subset of available visuals. To ground our study in representative practical use cases, we scope our task and dataset to daily-life, educational, and child-oriented narratives—domains that represent some of the most prevalent content creation scenarios where professional storytellers frequently operate. We further constrain the length of the generated narratives and the number of associated images to align with the conventions of typical children's

books and educational materials, ensuring that the resulting multimodal stories are directly usable in real-world contexts.

3.2 Dataset Creation

Inspired by the widespread adoption of synthetically generated datasets in recent work (Ouyang, 2025; Tan et al., 2025; Baes et al., 2025; Peper et al., 2025; Cao et al., 2025; Peng et al., 2025), we construct a new dataset tailored for this task through a multi-stage generation process involving both LLM-based planning and VLM-based image generation. The workflow proceeds as follows:

1. **Topic Brainstorming:** We use GPT-4.1 to generate 40 diverse story topics, each defined by a specific setting ("where") and main characters ("who").
2. **User Query Creation:** For each topic, GPT-4.1 generates 5 distinct, user-like textual queries that simulate how ordinary users might naturally request a story within that topic.
3. **Event Brainstorming:** For each user query, GPT-4.1 generates 20 distinct narrative events described in text. These events represent key moments that could occur within a story and serve as candidates for image generation.
4. **Image Generation:** Each textual event is used as input to GPT Image 1 (OpenAI, 2025) for image generation. For quality control, we conduct manual inspections verifying that GPT Image 1 consistently produces faithful event-grounded visuals.

This process results in a dataset with the following characteristics: (1) 40 topics and 5 user queries per topic, with **200 user queries in total**; (2) 100 images per topic, with **4,000 images in total**. The prompts we use are presented in Appendix A.

4 Method

In this section, we introduce our method for generating multimodal stories from image collections: *Narrative-Centric Image Selection and Content Planning*. Unlike prior work on multimodal storytelling (Tian et al., 2024a; Yang et al., 2024; An et al., 2024) which often assumes either unrestricted visual resources or fully open-ended narrative topics, our problem setting imposes a critical constraint: the available visual resources are finite, and the story must therefore be generated in direct correspondence with a carefully selected subset of the limited image collection. This setup closely reflects a common scenario faced by real-world content creators, where the story to be crafted is constrained to unfold entirely within the scope of a pre-existing image collection. We identify two key challenges that consistently emerge in practice: (1) *image selection*: how to retrieve a subset of images that are both thematically relevant and narratively useful from a large image collection; (2) *content planning*: how to construct an engaging and coherent story that effectively leverages the selected images. To address these challenges, we design a method that guides **image selection** (§4.1) and **content planning** (§4.2) from a **narrative-centric** perspective.

Intuitively, not all images in a collection contribute equally to storytelling. A good narrative typically progresses through a sequence of distinct yet connected events, with each advancing the plot in a meaningful way. Inspired by this observation, our narrative-centric image selection strategy aims to identify a subset of images that maximizes event diversity while remaining semantically and logically compatible within a narrative. Event diversity ensures that each selected image adds a unique piece to the story, which reduces redundancy and enriches the overall plot. At the same time, we avoid selecting images that are too semantically distant or logically inconsistent, as such choices would force the LLM to "invent" implausible connections. By balancing these two factors, we create a scaffold that provides both breadth and cohesion for subsequent multimodal story generation.

Once a meaningful subset of images has been selected, the next challenge is to transform these unordered images into a coherent narrative. To this end, we analyze the temporal cues and causal dependencies between events depicted in the images to infer a plausible ordering. This step plays an important role in improving story naturalness by ensuring that events unfold in an order that aligns with human expectations. Finally, recognizing that even a well-ordered sequence of images may still contain implicit narrative gaps, we explicitly bridge these gaps by generating intermediate events, thereby preserving logical flow and narrative continuity. This narrative planning-driven approach ensures that the resulting story reads as a cohesive whole rather than a disjointed sequence of image descriptions.

The step-by-step procedure of our method is detailed in subsection 4.1 and subsection 4.2.

4.1 Narrative-Centric Image Selection

Our method selects a subset of five images from a collection of 100 images associated with a topic. We first encode all images into a shared embedding space using CLIP (Radford et al., 2021). These embeddings are indexed in a vector store to support efficient similarity-based retrieval. Given a user query specifying a story intent, we retrieve the top-20 images most relevant to the query. For each of the 20 retrieved images, we prompt GPT-4.1 to generate a narrative-style event description capturing the main actions and situation depicted in the image. We then iteratively construct a set of five event descriptions maximizing narrative diversity while preserving semantic and logical compatibility:

1. *Initialization*: Among the 20 event descriptions, we prompt GPT-4.1 to identify the one that is most distinct from the rest. This event is added to the selected event pool, with the remaining events forming the unselected event pool.
2. *Iterative Selection*: We repeat the following steps until five events are selected:
   (a) **Diversity Ranking.** Given the current selected pool and unselected pool, we prompt GPT-4.1 to identify the event in the unselected pool that is most diverse relative to the events in the selected pool.
   (b) **Plausibility Check.** Before including this candidate event to the selected event pool, we prompt GPT-4.1 to determine whether it is semantically and logically compatible with the events that are already in the selected pool. If this event is deemed too semantically distant or logically inconsistent with any of the already selected events, it is discarded and the next most diverse candidate event is considered.

   This iterative selection process is designed to balance breadth and cohesion by ensuring that each selected event contributes novel narrative content, and that the resulting set of events can be reasonably connected into a story.

After four iterations, the selected event pool contains five event descriptions that collectively provide a diverse yet narratively grounded scaffold. These five events along with their corresponding images are returned as the final subset used for subsequent multimodal story generation. The prompts we use are presented in Appendix B.

4.2 Narrative-Centric Content Planning

With the subset of five selected images and their corresponding event descriptions obtained from narrative-centric image selection, our next goal is to transform these unordered events into a full story. We achieve this in three stages: (1) Ordering Events via Temporal and Causal Dependency Analysis, (2) Bridging Narrative Gaps with Intermediate Events, and (3) Expanding Story from Narrative Outline.

Ordering Events via Temporal and Causal Dependency Analysis. We first prompt GPT-4.1 with all five event descriptions to reason about their potential ordering from a temporal and causal perspective. Specifically, GPT-4.1 is asked to identify two types of pairwise event relations:

- *Strict precedence constraints*: pairs of events in which one must precede the other due to strong temporal or causal dependencies. GPT-4.1 is guided to identify these relations by examining temporal context cues (e.g., "early in the morning" or "after the class") and commonsense preconditions (e.g., "the lock must release before the door can open" or "the train must arrive before passengers can board"). Enforcing these constraints ensures that the resulting storyline is both temporally consistent and causally coherent, preventing implausible event progressions that would confuse readers or break narrative logic.
- *Soft ordering preferences*: pairs of events in which one certain ordering is not strictly required but is narratively preferable. Beyond strict dependencies, many event pairs exhibit a preferred directionality that, while not mandatory, enhances the naturalness and readability of the story. We guide GPT-4.1 to identify such event pairs by considering narrative conventions, such as presenting problems before resolutions, or moving from broader context to specific detail. Together with the strict constraints, these soft preferences allow the resulting storyline to not only "make sense" but also "feel right", and therefore to be aligned more closely with how humans intuitively structure stories.

GPT-4.1 outputs a list of such constraints and preferences, along with the underlying rationale for each ordering. These constraints and preferences are then jointly considered in a second round of reasoning, where we prompt GPT-4.1 again to produce a final ordering of the five events that best satisfies the identified constraints and preferences.

Bridging Narrative Gaps with Intermediate Events. Even with a well-ordered event sequence, the resulting storyline may still exhibit implicit gaps if important intermediate steps between consecutive events are missing, such as missing transitions, motivations, or actions. To address this, we prompt GPT-4.1 to examine the ordered event sequence, and then to generate novel intermediate events that function as bridging steps between the original five while staying faithful to the original user intent. These supplementary events serve to provide a smoother narrative flow and ensure that each segment of the story carries comparable narrative weight while preventing abrupt shifts in pacing or focus.

Expanding Story from Narrative Outline. Finally, we combine the ordered original events with any newly generated intermediate bridging events to construct a complete narrative outline. This enriched narrative outline is then provided to GPT-4.1 as input for the final expansion stage, where GPT-4.1 elaborates the narrative outline into a full story that faithfully reflects the original user intent. The prompts we use are presented in Appendix C. An example of multimodal stories generated by our method is presented in Figure 2.

5 Evaluation

We conduct four sets of experiments to comprehensively evaluate the effectiveness of our narrative-centric image selection and content planning method. In subsection 5.1, we perform automated evaluation using GPT-5 to rate the multimodal stories produced by our method and a baseline method across multiple evaluation criteria. In subsection 5.2, we perform human evaluation in which annotators make pairwise comparisons between stories generated by two methods and indicate their preferences with respect to multiple dimensions. In subsection 5.3, we perform ablation studies to verify the necessity and contribution of each component in our framework. In subsection 5.4, we perform fine-grained analyses of individual operations in our framework to examine whether they behave as intended.

5.1 Automated Evaluation

LLM-as-a-Judge has been frequently used for automatic assessment across a wide range of natural language generation tasks (Gilardi et al., 2023; Zheng et al., 2023; Liu et al., 2023b). Following this paradigm, we employ GPT-5 to automatically evaluate the quality of the multimodal stories produced by our method. For comparison, we construct a baseline method that retrieves the top-5 images most relevant to the user query, prompts GPT-4.1 to generate narrative-style event descriptions for each retrieved image, and then directly composes a full story from the five resulting events with GPT-4.1. To ensure a fair comparison, we reuse parts of the prompts introduced in Appendix B and Appendix C. Additionally, we control for output length in both methods by explicitly instructing GPT-4.1 to generate stories containing no more than 1,000 words.

For both methods, we evaluate a total of 200 multimodal stories, each generated in response to a unique user query across 40 distinct topics (5 queries per topic). Each sample consists of a natural language story accompanied by five associated images. We design a comprehensive evaluation protocol spanning multiple dimensions: the first three focus on single-modality aspects (text or image alone), while the remaining four capture cross-modal properties that jointly assess the interplay between text and images.

- **Logical Sense:** How well do the story's events follow plausible reasoning and causal structure?
- **Naturalness:** How fluent and human-like does the story feel in terms of storytelling style?
- **Visual Engagement:** How compelling, stimulating, and attention-grabbing is the visual component of the story?
- **Coherence:** How well do the text and images work together to form a unified continuous story?
- **Novelty:** How original and creative are the story's textual and visual components?
- **Text-Image Alignment:** How well do the text and images match in terms of content, semantics, and narrative focus?
- **Faithfulness to User Intent:** How well do the story's textual and visual components jointly fulfill the user's intended theme, tone, and narrative goal?

Following the standard practice in recent work on evaluating interleaved image–text generation (An et al., 2024; Liu et al., 2024b; Xu et al., 2025; Nie et al., 2025), for each multimodal story sample, we prompt GPT-5 to provide a numerical rating (1–5 scale) for every evaluation dimension based on the fine-grained rubrics defined in Appendix D.

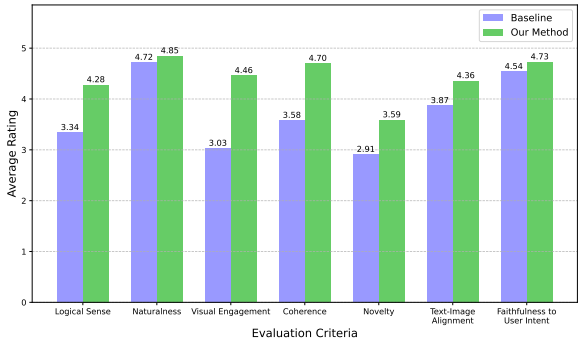

Figure 1: Automated evaluation results comparing our method with the baseline, showing the average GPT-5 ratings across seven evaluation dimensions.

Results are presented in Figure 1, where we report the average ratings across 200 samples for each evaluation dimension. Our method consistently outperforms the baseline by a substantial margin in logical sense, visual engagement, coherence, novelty, and text-image alignment, clearly demonstrating the effectiveness of our method. It is also observed that the performance improvements in naturalness and faithfulness to user intent are relatively modest. We further investigate this in the next section, where we conduct human evaluation to obtain a more accurate and nuanced assessment.

5.2 Human Evaluation

In addition to the automated evaluation conducted with the GPT-5 judge, we perform complementary human evaluation to provide a more comprehensive performance assessment. Human judgment has long been regarded as the gold standard for evaluating language generation tasks such as story generation (Celikyilmaz et al., 2020; Caglayan et al., 2020; van der Lee et al., 2021). Following this standard practice, we randomly sample 100 pairs of multimodal stories generated by our method and the baseline method in response to the same user queries. We then recruit human annotators to examine each pair and indicate their preferences through pairwise comparisons. To ensure annotation quality, we recruit annotators on Prolific ($12/hr) who speak English as their first language, and qualify them by first asking them a screening question then manually verifying their answers to discard annotations from those who fail the screening. Each story pair is evaluated by ten different annotators across the seven evaluation dimensions defined in subsection 5.1. To avoid potential positional bias, we randomize the presentation order of the two stories in each pair.

Results are presented in Table 1, which reports the percentage of times multimodal stories generated by each method are preferred across seven evaluation dimensions. Annotators' preferences with a fair average inter-annotator agreement measured by Fleiss's kappa (Fleiss, 1971) indicate that compared to the baseline, our method consistently produces more natural, coherent, and novel multimodal stories that logically make more sense while exhibiting better visual engagement and text-image alignment. The faithfulness to user intent remains comparable between the two methods, which we attribute to the reuse of prompts that instruct GPT-4.1 to reflect the original user intent during generation. We present a pair of multimodal story examples in Appendix E with a detailed comparison, where the one generated by our method (Figure 2) is preferred over the baseline output (Figure 3) across all seven evaluation dimensions, according to the majority vote of ten annotators.

5.3 Ablation Study

We conduct ablation studies to examine the contribution of the two key components of our method—narrative-centric image selection (§4.1) and content planning (§4.2)—to the overall quality of the generated multimodal stories. These ablation studies aim to both quantify their impact and verify their necessity by comparing the full method against ablated variants. Specifically, Ablation #1 retains narrative-centric content planning but skips image selection, while Ablation #2 performs narrative-centric image selection without content planning. We evaluate both ablated variants through human evaluation following the same setup described in subsection 5.2. Annotators' preferences, reported in Table 1 with a fair average inter-annotator agreement, indicate that both components are essential for enhancing the logical sense, naturalness, coherence, and novelty of multimodal stories. Moreover, we find that narrative-centric content planning, rather than image selection, plays a more critical role in improving text-image alignment. Finally, since visual engagement solely depends on the visual modality, we do not conduct ablations on this dimension, as narrative-centric content planning has no impact on it.

5.4 Analysis

Since the two components of our method comprise five core individual operations, it is necessary to verify the effectiveness of each operation through fine-grained analyses. For each operation, we ran-

Methods	Logical Sense			Naturalness			Visual Engagement		
	Win%	Lose%	Tie%	Win%	Lose%	Tie%	Win%	Lose%	Tie%
Ours v.s. Baseline	**82.7****	6.8	10.5	**51.4***	19.2	29.4	**48.0***	20.7	31.3
Ours v.s. Ablation #1	**73.9****	5.4	20.7	**38.8**	26.3	34.9	—	—	—
Ablation #1 v.s. Baseline	**78.4****	15.6	6.0	**45.3**	21.4	33.3	—	—	—
Ours v.s. Ablation #2	**76.2****	15.0	8.8	**51.0***	16.1	32.9	—	—	—
Ablation #2 v.s. Baseline	**69.4****	8.3	22.3	**43.7**	19.0	37.3	—	—	—

Methods	Coherence			Novelty			Text-Image Alignment			Faithfulness to User Intent		
	Win%	Lose%	Tie%	Win%	Lose%	Tie%	Win%	Lose%	Tie%	Win%	Lose%	Tie%
Ours v.s. Baseline	**86.5****	3.9	9.6	**74.0****	15.3	10.7	**41.3**	22.5	36.2	20.1	13.0	**66.9****
Ours v.s. Ablation #1	**70.3****	16.4	13.3	**58.6****	18.0	23.4	29.7	21.6	**48.7***	28.6	10.2	**61.2****
Ablation #1 v.s. Baseline	**81.0****	13.1	5.9	**49.7***	22.8	27.5	**44.2**	19.4	36.4	20.7	8.5	**70.8****
Ours v.s. Ablation #2	**75.9****	17.5	6.6	**52.1***	14.5	33.4	**49.5***	25.4	25.1	14.3	12.8	**72.9****
Ablation #2 v.s. Baseline	**63.8****	15.3	20.9	**66.5****	6.4	27.1	26.2	32.2	**41.6**	15.0	17.7	**67.3****

Table 1: Human evaluation results for both the comparison against the baseline and the ablation studies, showing the percentage of times multimodal stories generated by each method are preferred across seven evaluation dimensions. "Win" means the method on the left is preferred. ** indicates results are significant at $p < 0.01$ (* at $p < 0.05$) confidence level. We do not conduct ablations on visual engagement, as narrative-centric content planning has no impact on this single-modal property.

domly sample 100 instances and recruit 10 annotators to examine them and assess whether the operation behaves as intended. Results are as follows:

- **Diversity Ranking:** In 94% of cases, each of the selected five events adds new narrative content, characters, settings, actions, or plot directions not covered by the rest four.
- **Plausibility Check:** In 90% of cases, each of the selected five events is semantically and logically compatible with the rest four.
- **Ordering Events via Temporal and Causal Dependency Analysis:** In 88% of cases, the identified constraints/preferences are reasonable and the final ordering of five events plausibly satisfies them.
- **Bridging Narrative Gaps with Intermediate Events:** In 95% of cases, the generated intermediate events effectively function as bridging steps between the original five while staying faithful to the original user intent and avoiding any logical flaws or contradictions.
- **Expanding Story from Narrative Outline:** In 97% of cases, the narrative outline is expanded into a coherent and engaging story which maintains a clear arc, smooth transitions and natural pacing while staying faithful to the original user intent.

These results demonstrate that the five individual operations function reliably in isolation, and that errors do not propagate or amplify downstream: the high per-step accuracy across all five operations indicates that the framework is robust against early-stage imperfections.

6 Conclusions

We presented a new task, multimodal story generation from image collections, which departs from traditional multimodal storytelling paradigms by introducing realistic resource constraints into the generation process. This setting captures a widely encountered yet underexplored scenario in real-world content creation, where stories must be composed using a limited subset of images that must be retrieved from pre-existing visual assets. To address the inherent challenges of this problem, we proposed a narrative-centric image selection and content planning framework: it retrieves a diverse and narratively compatible subset of images, organizes them into a coherent event sequence, bridges narrative gaps between them, and finally expands into a full story. Our approach consistently outperforms baseline methods across a range of evaluation criteria, as verified by both automated metrics and human judgments. Beyond demonstrating the effectiveness of our method, we also contribute the first dataset tailored for this new task, enabling future research on multimodal story generation under realistic resource constraints.

7 Limitations

While our work represents an initial step toward multimodal story generation from image collections, two limitations remain.

First, our study focuses on daily-life, educational, and child-oriented narratives, which, while representative of many real-world use cases, does not capture the full diversity of storytelling scenarios. Future work could extend the task to more domains such as long-form fiction, news narratives, or multimodal documentaries.

Second, our study is currently limited to the English language. All datasets, prompts, and evaluations are designed and conducted in English, which means that our findings may not generalize to multilingual or cross-lingual storytelling settings. Future work could extend the task to multilingual datasets and explore how narrative-centric image selection and content planning behave in different linguistic settings.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Jie An, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Lijuan Wang, and Jiebo Luo. 2024. Openleaf: A novel benchmark for open-domain interleaved image-text generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11137–11145.

Naomi Baes, Raphael Merx, Nick Haslam, Ekaterina Vylomova, and Haim Dubossarsky. 2025. LSC-eval: A general framework to evaluate methods for assessing dimensions of lexical semantic change using LLM-generated synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10905–10939, Vienna, Austria. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2020. Curious case of language generation evaluation metrics: A cautionary tale. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2322–2328, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shuyang Cao, Kaijian Zou, and Lu Wang. 2025. SYNC: A synthetic long-context understanding benchmark for controlled comparisons of model capabilities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33615–33636, Suzhou, China. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Dongping Chen, Ruoxi Chen, Shu Pu, Zhaoyi Liu, Yanru Wu, Caixi Chen, Benlin Liu, Yue Huang, Yao Wan, Pan Zhou, and Ranjay Krishna. 2025a. Interleaved scene graphs for interleaved text-and-image generation assessment. In *The Thirteenth International Conference on Learning Representations*.

Hong Chen, Rujun Han, Te-Lin Wu, Hideki Nakayama, and Nanyun Peng. 2022a. Character-centric story visualization via visual planning and token alignment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8259–8272, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hong Chen, Yifei Huang, Hiroya Takamura, and Hideki Nakayama. 2021. Commonsense knowledge aware concept selection for diverse and informative visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 999–1008.

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

Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. 2025b. Comm: A coherent interleaved image-text dataset for multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8073–8082.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022b. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elizabeth Clark, Yangfeng Ji, and Noah A. Smith. 2018. Neural text generation in stories using entity representations as context. In *Proceedings of the 2018 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2250–2260, New Orleans, Louisiana. Association for Computational Linguistics.

Kenneth D Crews and Melissa A Brown. 2011. *Copyright, Museums, and Licensing of Art Images*. Citeseer.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660, Florence, Italy. Association for Computational Linguistics.

Chun-Mei Feng, Yang Bai, Tao Luo, Zhen Li, Salman Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Yong Liu. 2025. Vqa4cir: Boosting composed image retrieval with visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2942–2950.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and 1 others. 2023. Talecrafter: Interactive story visualization with multiple characters. *arXiv preprint arXiv:2305.18247*.

Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual writing prompts: Character-grounded story generation with curated image sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581.

Chi-yang Hsu, Yun-Wei Chu, Ting-Hao Huang, and Lun-Wei Ku. 2021. Plot and rework: Modeling storylines for visual storytelling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4443–4453, Online. Association for Computational Linguistics.

Ting-Yao Hsu, Chieh-Yang Huang, Yen-Chia Hsu, and Ting-Hao Huang. 2019. Visual story post-editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6581–6586, Florence, Italy. Association for Computational Linguistics.

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

Qiuyuan Huang, Zhe Gan, Asli Celikyilmaz, Dapeng Wu, Jianfeng Wang, and Xiaodong He. 2019. Hierarchically structured reinforcement learning for topically coherent visual story generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8465–8472.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*, pages 17283–17300. PMLR.

Siqi Kou, Jiachun Jin, Zhihong Liu, Chang Liu, Ye Ma, Jian Jia, Quan Chen, Peng Jiang, and Zhijie Deng. 2025. Orthus: Autoregressive interleaved image-text generation with modality-specific heads. In *Forty-second International Conference on Machine Learning*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. MM-EMBED: UNIVERSAL MULTIMODAL RETRIEVAL WITH MULTIMODAL LLMS. In *The Thirteenth International Conference on Learning Representations*.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. 2024a. Intelligent grimm - open-ended visual storytelling via latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6190–6200.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. 2024b. Holistic evaluation for interleaved text-and-image generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22002–22016, Miami, Florida, USA. Association for Computational Linguistics.

Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134.

Adyasha Maharana and Mohit Bansal. 2021. Integrating visuospatial, linguistic, and commonsense structure into story visualization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6772–6786, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2021. Improving generation and evaluation of visual stories via semantic consistency. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2427–2442, Online. Association for Computational Linguistics.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2022. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European conference on computer vision*, pages 70–87. Springer.

Ming Nie, Chunwei Wang, Jianhua Han, Hang Xu, and Li Zhang. 2025. Towards unified multimodal interleaved generation via group relative policy optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

OpenAI. 2025. Gpt-image-1. `https://platform.openai.com/docs/models/gpt-image-1`.

Jialin Ouyang. 2025. TreeCut: A synthetic unanswerable math word problem dataset for LLM hallucination evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1073–1085, Vienna, Austria. Association for Computational Linguistics.

Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. 2024. Synthesizing coherent story with auto-regressive latent diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2920–2930.

Bo Peng, Zhiheng Wang, Heyang Gong, and Chaochao Lu. 2025. IP-dialog: Evaluating implicit personalization in dialogue systems with synthetic data. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17007–17040, Suzhou, China. Association for Computational Linguistics.

Joseph J Peper, Wenzhao Qiu, Ali Payani, and Lu Wang. 2025. MDBench: A synthetic multi-document reasoning benchmark generated with knowledge guidance. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25592–25621, Vienna, Austria. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. 2023. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2493–2502.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Fei Shen, Hu Ye, Sibo Liu, Jun Zhang, Cong Wang, Xiao Han, and Yang Wei. 2025. Boosting consistency in story visualization with rich-contextual conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6785–6794.

Xiaoqian Shen and Mohamed Elhoseiny. 2025. Storygpt-v: Large language models as consistent story visualizers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13273–13283.

Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. 2021. Instance-level image retrieval using reranking transformers. In *Proceedings of the IEEE/CVF Interna-*

tional Conference on Computer Vision (ICCV), pages 12105–12115.

Juntao Tan, Liangwei Yang, Zuxin Liu, Zhiwei Liu, Rithesh R N, Tulika Manoj Awalgaonkar, Jianguo Zhang, Weiran Yao, Ming Zhu, Shirley Kokane, Silvio Savarese, Huan Wang, Caiming Xiong, and Shelby Heinecke. 2025. PersonaBench: Evaluating AI models on understanding personal information through accessing (synthetic) private user data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 878–893, Vienna, Austria. Association for Computational Linguistics.

Giulia Taurino, Sarah Sweeney, Drew Facklam, and David A Smith. 2025. Copyright, privacy, and public access in news archives: a proof of concept on the boston globe photograph morgue. *AI & SOCIETY*, pages 1–19.

Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, and 1 others. 2024a. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024b. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Jing Wang, Jianlong Fu, Jinhui Tang, Zechao Li, and Tao Mei. 2018a. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ruize Wang, Zhongyu Wei, Piji Li, Qi Zhang, and Xuanjing Huang. 2020. Storytelling from an image stream using scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9185–9192.

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018b. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 899–909, Melbourne, Australia. Association for Computational Linguistics.

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 *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3022–3029.

Zhiyang Xu, Minqian Liu, Ying Shen, Joy Rimchala, Jiaxin Zhang, Qifan Wang, Yu Cheng, and Lifu Huang. 2025. Modality-specialized synergizers for interleaved vision-language generalists. In *The Thirteenth International Conference on Learning Representations*.

Jheng-Hong Yang, Carlos Lassance, Rafael Sampaio De Rezende, Krishna Srinivasan, Miriam Redi, Stéphane Clinchant, and Jimmy Lin. 2023a. Atomic: An image/text retrieval test collection to support multimedia content creation. In *Proceedings of the 46th International ACM SIGIR conference on research and development in information retrieval*, pages 2975–2984.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023b. DOC: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3378–3465, Toronto, Canada. Association for Computational Linguistics.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. 2024. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Licheng Yu, Mohit Bansal, and Tamara Berg. 2017. Hierarchically-attentive RNN for album summarization and storytelling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 966–971, Copenhagen, Denmark. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, Lirui Zhao, Shuo Liu, Tianhua Li, Yuxuan Xie, Xiaojun Chang, Yu Qiao, Wenqi Shao, and Kaipeng Zhang. 2025. Opening: A

comprehensive benchmark for judging open-ended interleaved image-text generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 56–66.

Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. 2023. R2former: Unified retrieval and reranking transformer for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19370–19380.

A Prompts for Dataset Creation

Topic Brainstorming

```
You are an expert story designer tasked with brainstorming creative and diverse story topics
for a multimodal storytelling dataset. Each topic must be suitable for children's books,
educational content, or daily-life storytelling scenarios.

Generate **40 unique story topics**. Each topic must include:
- A specific setting (WHERE the story takes place)
- Main characters (WHO is involved in the story)

Guidelines:
- Main characters in every story must include a young kid named John. You may include other
supporting characters if relevant, but John should always be the central character.
- Make the settings and characters highly diverse across all examples.
- Avoid repetition or generic placeholders.
- Return the results as a Python list.
```

User Query Creation

```
You are given a short story background consisting of two elements:
- **Where** the story takes place: $WHERE.
- **Who** is involved in the story: $WHO.

Your task is to generate **5 distinct user queries** that an ordinary person might type when
asking an AI assistant to create a story based on this background.

Guidelines:
- Keep queries short, clear, and easy to understand.
- Make queries sound natural, like real user requests (not overly formal instructions).
- The 5 queries should reflect a variety of possible user intentions, including but not limited
to:
  - requesting setting details
  - focusing on character interactions
  - asking for plot twists or conflicts
- Ensure diversity across the 5 queries.
- Return the results as a Python list.
```

Event Brainstorming

```
You are given three inputs that define the story request and background (where and who):

* **Story Request**: $USER-QUERY
* **Where** the story takes place: $WHERE.
* **Who** is involved in the story: $WHO.

Based on the provided story request and background, your task is to generate **four distinct
sets of event moments**, each set outlining a coherent narrative arc for a possible story. Each
set should contain **exactly five event moments**, expressed as **one concise sentence per
event moment**.

### Requirements

1. **Story coverage**:

   * Each set of five event moments should represent the **major components** (e.g., setup,
   escalation, turning point, climax, resolution) of a coherent story.
   * The full story implied by each set should naturally extend **beyond five event moments**,
   but the selected five should capture the **most important milestones**.

2. **Diversity**:

   * The four sets must be **different from one another**, in order to offer distinct narrative
   trajectories.
   * Within each set, all five event moments must also be **non-redundant** and cover different
   aspects of the story.

3. **Image-friendliness**:

   * Each event moment should be **visually concrete** and **easy to depict** in an image.
   * Avoid overly complex or abstract descriptions that would be difficult to illustrate.

4. **Form**:

   * Each event moment must be written as **exactly one sentence**.
   * Use simple, vivid, and narrative-friendly language.

### Output Format

Return the results strictly as a **Python list of lists**, where:

* The outer list contains four inner lists (one per story set).
* Each inner list contains exactly five strings (one sentence per event moment).
```

Image Generation

```
You are given an **event moment** expressed as a concise sentence.

* **Event moment**: $EVENT-MOMENT

Your task is to generate **one image** that illustrates the given event moment.

### Requirements

1. **The generated image must stay true to the semantics of the given event moment**:

   * **Text-to-visual alignment**: The central actions, characters, objects, and settings
   described in the event moment should be clearly reflected in the generated image.
   * **Detail enhancement**: It is acceptable to add fine-grained contextual details (e.g.,
   background decorations, incidental characters, props, environmental cues) as long as they
   enhance visual storytelling and do not contradict the event moment description.

2. **The generated image should adopt a Cartoon style**:

   * **Character design**: Characters should have rounded and child-friendly proportions.
   * **Color palette and line work**: Use bright, warm colors and clean outlines.
   * **Visual clarity**: Characters, objects, and environments must be distinct and
   unambiguous, avoiding clutter or confusing compositions.
   * **Misc.**: John is a young kid who wears glasses and a blue T-shirt (ignore this item if
   John is not mentioned in the event moment).
```

B Prompts for Narrative-Centric Image Selection

Event Description Generation

```
You are given an image depicting a real-world scene. Your task is to generate a concise yet
informative **narrative-style event description** that captures the **main actions** and
**situation** depicted in the image.

**Guidelines:**
- Focus on the main event and describe the key action(s) taking place, rather than listing all
objects or background details.
- Identify the key participant(s) and refer to them naturally within the narrative context. In
particular, if one of the participants in the image is John (a young kid wearing glasses and a
blue T-shirt), refer to him by name.
- Go beyond a literal caption and include essential contextual cues (e.g., setting, intent,
mood, or interaction) that help a reader understand the situation without seeing the image.
- Write in a coherent story-like tone as if this event description were part of a larger
narrative.

**Output Format:** A short narrative-style event description (1–2 sentences).
```

Initialization

```
You are given a list of narrative-style event descriptions, each summarizing the main actions
and situation depicted in a different image related to the same topic.

**EVENT LIST:**
$EVENT-LIST

Your task is to identify the SINGLE event that is most distinct from all the others in terms of
its narrative content, setting, characters, and action.
- "Distinct" means the event introduces a storyline element that is clearly different from the
majority, not just a small detail variation.
- Avoid events that are too similar to others or repeat common actions.
- Focus on those that offer a new narrative perspective or plot direction.

**Output Format:**
- Index: [0-based index of the most distinct event in EVENT LIST]
```

Diversity Ranking

```
You are given two sets of narrative event descriptions:

**SELECTED POOL:**
$SELECTED-POOL

**UNSELECTED POOL:**
$UNSELECTED-POOL

Your task is to identify the event in the Unselected Pool that is most diverse relative to the
events in the Selected Pool.
- "Diverse" means the event adds NEW narrative content, characters, settings, actions, or plot
directions not already covered by the Selected Pool.
- Prioritize events that introduce COMPLEMENTARY story elements rather than redundant ones.

**Output Format:**
- Index: [0-based index of the most diverse event in UNSELECTED POOL]
```

Plausibility Check

```
You are given:

**SELECTED POOL:**
$SELECTED-POOL

**CANDIDATE EVENT:**
$CANDIDATE-EVENT

Your task is to evaluate whether the Candidate Event is both SEMANTICALLY and LOGICALLY
compatible with the events in the Selected Pool.
- "Semantically compatible" means the characters, setting, and world are consistent and
plausible within the existing context.
- "Logically compatible" means the event could plausibly occur before, after, or alongside the
others without contradicting them.
- If the Candidate Event is both SEMANTICALLY and LOGICALLY compatible with all the events in
the Selected Pool, return "Compatible"; if the Candidate Event is NOT SEMANTICALLY or LOGICALLY
compatible with at least one of the events in the Selected Pool, return "Not compatible".

Here are a few examples with explanations to guide your reasoning:
$FEW-SHOT-EXAMPLES

**Output Format:**
- Verdict: "Compatible" or "Not compatible"
```

C Prompts for Narrative-Centric Content Planning

Pairwise Event Relation Identification

```
You are a narrative reasoning expert analyzing temporal/causal relations.

You are given 5 unordered narrative event descriptions:

**EVENT LIST:**
$EVENT-LIST

**Task:** Output two lists:
1) Strict Precedence Constraints: pairs where one must occur before the other, based on
temporal context cues and commonsense preconditions.
2) Soft Ordering Preferences: pairs where a direction is preferred but not mandatory, based on
narrative conventions, natural progression, and readability.

For each pair, explain WHY in 1–2 sentences.

Here are a few examples with explanations to guide your reasoning:
$FEW-SHOT-EXAMPLES

Return ONLY a Python dictionary following this schema:
{
  "strict_constraints": [
    {"before": "EVENT_X1", "after": "EVENT_Y1", "reason": "..."},
    {"before": "EVENT_X2", "after": "EVENT_Y2", "reason": "..."},
    ...
  ],
  "soft_preferences": [
    {"preferred_before": "EVENT_A1", "preferred_after": "EVENT_B1", "reason": "..."},
    {"preferred_before": "EVENT_A2", "preferred_after": "EVENT_B2", "reason": "..."},
    ...
  ]
}
```

Final Ordering Judgment

```
You are an expert story planner. Create the most plausible/natural order that satisfies ALL
strict constraints and as many soft preferences as possible.

**EVENT LIST:**
$EVENT-LIST

**STRICT CONSTRAINTS:**
$STRICT-CONSTRAINTS

**SOFT PREFERENCES:**
$SOFT-PREFERENCES

Return ONLY a Python dictionary following this schema:
{
  "ordered_events": ["EVENT_1", "EVENT_2", "EVENT_3", "EVENT_4", "EVENT_5"]
}
```

Bridging Narrative Gaps with Intermediate Events

```
You are a professional story writer. Examine each consecutive event pair in ORDERED EVENTS and,
ONLY if a meaningful gap exists (missing transition/motivation/action), add ONE novel bridging
event. Make sure to stay faithful to the user query indicating a story intent.

**USER QUERY:**
$USER-QUERY

**ORDERED EVENTS:**
$ORDERED-EVENTS

**Guidelines:**
- Add a bridging event only when warranted; skip event pairs that already read smoothly.
- Keep each bridging event 1–2 sentences and ensure temporal/causal plausibility.

Return ONLY a Python dictionary following this schema:
{
  "bridging_events": [
    {"between": ["EVENT_i", "EVENT_{i+1}"], "bridge_event": "..."},
    {"between": ["EVENT_j", "EVENT_{j+1}"], "bridge_event": "..."},
    ...
  ]
}
```

Expanding Story from Narrative Outline

```
You are a professional story writer. Expand the narrative outline into a coherent, engaging
story (no more than 1000 words). Maintain a clear arc, smooth transitions, and natural pacing.
Make sure to stay faithful to the user query indicating a story intent.

**USER QUERY:**
$USER-QUERY

**NARRATIVE OUTLINE:**
$NARRATIVE-OUTLINE
```

D Evaluation Rubrics

- **Logical Sense:** How well do the story's events follow plausible reasoning and causal structure?
  - 5: All events follow plausible reasoning and causal structure.
  - 4: Most events are logically connected and plausible, with only minor inconsistencies or unclear causal links.
  - 3: The story is mostly understandable but contains a few questionable or unexplained event transitions.
  - 2: Logical flaws, contradictions, or gaps frequently occur, making the story difficult to follow.
  - 1: The story lacks internal logic, with gaps or contradictory events that break narrative reasoning.

- **Naturalness:** How fluent and human-like does the story feel in terms of storytelling style?
  - 5: The story reads naturally, resembling a human-written narrative with realistic language.
  - 4: The story reads mostly naturally, with only occasional awkward phrasing.
  - 3: The story is readable but contains multiple unnatural phrases or stiff expressions.
  - 2: The story contains frequent awkward phrasing that disrupts immersion and readability.
  - 1: The story is highly unnatural or machine-like, with very poor readability.

- **Visual Engagement:** How compelling, stimulating, and attention-grabbing is the visual component of the story?
  - 5: The images are visually rich and emotionally evocative, greatly enhancing immersion and storytelling impact.
  - 4: The images are engaging and relevant, though some could be more impactful or diverse.
  - 3: The images are moderately engaging but feel repetitive or lack strong narrative momentum.
  - 2: The images are bland, uninspired, or fail to enhance the story's atmosphere.
  - 1: The images are uninteresting, irrelevant, or detract from the narrative experience.

- **Coherence:** How well do the text and images work together to form a unified continuous story?
  - 5: Text and images interweave seamlessly into a single continuous narrative. Visuals build cohesively on the written story and contribute to a clear unfolding progression from beginning to end.
  - 4: The story generally flows smoothly across text and images, with only minor disruptions in continuity.
  - 3: The overall storyline is understandable, but noticeable breaks, uneven pacing, or weak transitions between textual and visual parts reduce the sense of a cohesive whole.
  - 2: Frequent disruptions, inconsistencies, or abrupt narrative jumps between text and images make the story feel fragmented or loosely stitched together.
  - 1: Text and images fail to come together as a single narrative. The story feels disjointed, disorganized, or difficult to follow as a continuous multimodal sequence.

- **Novelty:** How original and creative are the story's textual and visual components?
  - 5: The story is highly original, imaginative, and surprising, offering unique perspectives or creative elements.
  - 4: The story is generally creative with some fresh elements, though parts feel familiar.
  - 3: The story contains a mix of predictable and moderately novel aspects.
  - 2: The story is mostly generic or derivative, with little originality.
  - 1: The story is completely conventional, cliché-ridden, or repetitive with no creative value.

- **Text-Image Alignment:** How well do the text and images match in terms of content, semantics, and narrative focus?
  - 5: All images accurately depict and enrich their corresponding textual events, creating a tightly integrated multimodal narrative.
  - 4: Most images align well with the text, with only minor mismatches or ambiguities.
  - 3: Several images are loosely related to the text but fail to deepen the story's meaning.
  - 2: Frequent misaligned or irrelevant visuals disrupt the connection between text and images.
  - 1: The images are largely unrelated to the textual narrative.
- **Faithfulness to User Intent:** How well do the story's textual and visual components jointly fulfill the user's intended theme, tone, and narrative goal?
  - 5: The story is fully faithful to the desired theme, tone, and narrative goal.
  - 4: The story is largely faithful with minor deviations or missing nuances.
  - 3: The story is partially faithful to the user intent but misses some aspects.
  - 2: The story contains frequent divergence from the user intent.
  - 1: The story ignores or contradicts the user's intended narrative direction.

E A Pair of Multimodal Story Examples

We present a pair of multimodal story examples in Figure 2 and Figure 3, where our method is preferred over the baseline across seven evaluation dimensions according to human annotators' majority votes. Compared with the baseline output, the multimodal story generated by our method demonstrates a higher level of narrative quality and sophistication. The events in our story unfold in a logically coherent sequence, with each development naturally leading to the next rather than feeling arbitrarily stitched together. This strong causal and temporal structure enables the narrative to follow a clear and compelling dramatic arc—beginning with a well-established exposition, steadily building tension through rising action, reaching a climactic turning point, and then resolving gracefully through falling action and denouement. As a result, our story reads more like a carefully crafted human-written narrative, complete with engaging twists and thoughtfully designed plot turns that keep the reader captivated from start to finish. In contrast, the baseline story suffers from loose and sometimes forced event connections, giving the impression of a disjointed list of happenings rather than a cohesive narrative. Its plot progression is flat and monotonous, lacking the emotional highs and lows that make a story memorable, and thus the overall reading experience is relatively dull. Furthermore, our story showcases greater diversity and richness in visuals and the resulting narrative scope—it spans a broader temporal range across two days and transitions from daytime to nighttime settings, adding depth and dynamism to the storytelling. The baseline, by comparison, remains confined to a shorter, single-day timeline, which limits narrative variety and makes the story feel more constrained and less immersive.

It was a crisp autumn afternoon when John arrived at Pinewood National Forest with his parents for a long-awaited camping trip. The leaves danced in shades of amber and crimson, and the air buzzed with a quiet excitement. John, a curious and observant ten-year-old who'd always dreamed of becoming a park ranger, felt as if this weekend had leaped straight out of his dreams. Shortly after setting up their tent near the visitor center, a ranger called the campers together for a safety briefing at the trailhead. A large bear warning sign stood beside the clearing, looming over the gathered guests. Together, families circled around the ranger as he spoke. "Bears are rarely aggressive," the ranger explained, his voice calm but firm, "but you need to know what to do if you encounter one. Always stay calm. Never run. Make yourself look tall. And most importantly, stay together." John stood close to the sign, his eyes wide, soaking in every word. He asked thoughtful questions: "What if the bear has cubs?" "What should we do if it comes near our campsite?" His earnestness made the ranger smile and nod at him in encouragement. Afterwards, the families scattered—some wandered to picnic sites, others set out on gentle trails. John and his parents decided on the Whispering Pines Loop, known for its sweeping views and glimpses of wildlife. The sun was setting by the time they returned to camp for dinner, and as the first stars appeared, a ranger jeep streaked past, siren lights flashing through the trees. Two rangers dove out and began searching the woods with flashlights, their beams cutting through the growing darkness. John watched in silence, unease twisting in his chest. The head ranger took notice and approached quietly. "There's been a report," he confided. "A family's child hasn't returned from Ridgeline Trail. We think their child may have gotten separated." Hearing this, John's heart pounded. Ridgeline Trail cut deeper into the forest with rougher terrain. He wanted desperately to help but knew he should stay back. Still, something told him his weekend in the forest would change by morning.

At dawn, John was awake before anyone else, sketchpad in hand, walking the near trails and stopping to draw clusters of pinecones and mushrooms. As he rounded a bend, something unusual caught his eye—a torn piece of red fabric, fluttering where it had snagged on a branch. A chill crept up his spine. He knelt to examine it and saw small, scattered footprints leading off the main trail. They looked like they belonged to a child, messy as if the owner had been running or limping. "Rangers!" John called, his voice clear and urgent. "Over here!" Within minutes, rangers hurried over, eyes scanning John's discovery. One used a loudspeaker to call out for the missing person, while another pulled out a compact emergency kit. "You did good, John," the head ranger said, voice steady. "She can't be far." They followed the path with caution, moving deeper into the brush. At last, behind a low stone outcrop near a trickling brook, they found her: a girl in a red shirt, sitting on the ground, face pale and streaked with tears. "I tried to find my parents," she whimpered. "But then a raccoon scared me, and I fell." Her knee was badly bruised, clearly twisted, and it was obvious she'd spent a cold, lonely night outdoors. "You're safe now," John assured her gently. "They're coming." He stayed at her side, comforting her while a ranger bandaged the injury, whispering encouragement and softly distracting her from the pain. "You were really brave," the ranger told the girl. "Just like John here." Moments later, the girl's parents—faces drawn with worry and hope—arrived, falling to their knees in relief. Tears and gratitude flowed freely as the head ranger radioed for an ATV to transport the little girl safely back.

The rest of the day blurred into warmth and celebration. John, modest but proud, was hailed as a hero at the visitor center. He spent the afternoon with the rangers, learning how to use a two-way radio, read trail maps, and spot animal tracks. To his surprise, the head ranger presented him with a shiny "Junior Ranger" badge. The rescued girl—smiling now, wrapped in a soft blanket—stood with her family, waving shyly at John. The two rangers strolled over, shaking John's hand. "You really have the heart of a ranger," the head ranger told him. "Maybe someday, we'll see you out here in uniform." John grinned wide, pride shining in his eyes. The other ranger added, "You didn't just help us—you brought a family back together. Thank you." At sunset, the two rangers waved goodbye to John, the girl, and her grateful family, then disappeared toward their jeep. That night, tucked in his sleeping bag beneath a canopy of moonlit branches, John realized the forest wasn't just a place of adventure and beauty—it was a place that asked for courage, vigilance, and heart. And deep inside, John knew that, one day, he wouldn't just be camping in these woods. He'd be protecting them. Just like a ranger.

Figure 2: A multimodal story example generated by our method, in response to the user query "Give me a story where John helps the park rangers solve a problem. (John is a young kid who wears glasses and a blue T-shirt)." We apply distinct color highlights to both the images and the story segments to visually illustrate their correspondence. [GPT-5 Judge's Ratings: Logical Sense: 5/5, Naturalness: 5/5, Visual Engagement: 5/5, Coherence: 5/5, Novelty: 4/5: Text-Image Alignment: 5/5, Faithfulness to User Intent: 5/5.]

On a gentle summer morning, John laced up his hiking boots and straightened his blue T-shirt, ready for another day at Pine River Park. Today was special. For weeks, he'd admired the work of the park rangers, always hoping for a chance to join in their adventures. As he adjusted his glasses, he spotted a familiar green truck rolling up near the forest trailhead. His heart skipped—a pair of rangers, badges shining, stepped out with a friendly wave. "Hey there! You must be John," called Ranger Val, her hands full of rescue gear and a walkie-talkie dangling at her side. The second ranger, Pete—tall, broad-shouldered, but with a kind smile—gave John a respectful nod. The rangers were clearly prepared for a busy day, and John couldn't believe his luck as he greeted them eagerly.

The trio quickly set off into the lush forest, sunlight streaming through the thick boughs overhead. John peppered the rangers with questions about animal habitats and safety: "How do you know if there's a bear nearby? What's the strangest thing you've ever found?" As they walked, Val and Pete explained everything and even let John carry a spare set of binoculars. Not long into their hike, John spotted a group of volunteers gathered on the main path, collecting litter with bags and gloves. Among them was a young girl balancing a stack of rusty cans and an older man straining to pick up a crumpled soda bottle in the underbrush. The rangers and John joined in, and, with lively conversation and plenty of teamwork, they cleared wrappers, bottles, and old newspapers from the trail. John's spirit was infectious—under his watchful eye, nobody overlooked even the smallest scrap of trash. By the time they finished, the path was pristine, and a real sense of pride buzzed in the afternoon air. After saying goodbye to the volunteers, the rangers led John deeper into the woods. They stopped near a sturdy wooden sign with a painted bear paw, where a small crowd of anxious visitors had gathered. The mood shifted as the group took in the serious warning: "BEAR ACTIVITY—STAY ALERT." Pete raised his hand to settle everyone down and began explaining what to do if they encountered wildlife. "Stay calm. Never run. Make yourself look big," he instructed, scanning the nervous faces. John listened with rapt attention, soaking up the practical advice. The importance of being prepared struck him—nature was incredible, but it demanded respect.

After the talk, Val suggested they check by the riverside. As they slipped through a patch of tall ferns, John heard a faint whimpering sound. At the muddy water's edge, a small otter sat hunched, twisting in distress. Its paw was tangled in a length of fishing line. Heart pounding, John knelt beside the creature, whispering soothing words while the rangers worked quickly. Val snipped the string away, and John gently comforted the otter, who soon slipped free and darted into the rushing water, safe and relieved. The teamwork filled him with a newfound confidence. They decided to circle back toward the ranger station, but as they approached the grassy clearing, John stopped suddenly. His sharp eyes caught something odd—deep paw prints, pressed fresh into the dirt, trailing off toward the trees. Excitedly, he pointed them out. The rangers hurried over, examining the tracks with curiosity and concern. "Good catch, John," Pete said, kneeling to study the prints. "These are too big for a dog… and see the shape? Could be a young bear wandering through." Just then, a mischievous raccoon crept out from behind John's backpack, rooting around for a snack and adding to the scene's lively commotion. With their adrenaline rising, the group quickly sprang into action. Val cordoned off the area and marked the tracks for further study, while Pete called in an update to the ranger base. They worked as a team, following the trail, making sure no visitors accidentally crossed paths with the wandering animal. The rangers praised John's sharp eyes—his discovery had helped them keep both the bear and the people safe.

Later, as the sun dipped behind the hills, John returned to the ranger station. He felt exhilarated but also humbled by the day's challenges and teamwork. The rangers thanked him warmly, shaking his hand as if he were one of their own. As he walked home, John thought of all he had learned—about nature's wonders and dangers, about teamwork, about how even small actions could make a difference. That evening, when his parents asked about his day, John beamed. "I got to help the rangers save an otter and keep the park safe. I think I found my calling," he declared, the excitement of a day well spent shining in his eyes. And so, John knew—he'd return to Pine River Park again and again, always ready to lend a helping hand.

Figure 3: A multimodal story example generated by baseline method, in response to the user query "Give me a story where John helps the park rangers solve a problem. (John is a young kid who wears glasses and a blue T-shirt)." We apply distinct color highlights to both the images and the story segments to visually illustrate their correspondence. [GPT-5 Judge's Ratings: Logical Sense: 4/5, Naturalness: 5/5, Visual Engagement: 4/5, Coherence: 5/5, Novelty: 4/5: Text-Image Alignment: 4/5, Faithfulness to User Intent: 5/5.]