SEED-Story: Multimodal Long Story Generation with Large Language Model

Shuai Yang [1*] Yuying Ge [2†] Yang Li [1] Yukang Chen [4] Yixiao Ge [2,3]
Ying Shan [2,3] Yingcong Chen [1,5†]

[1]HKUST(GZ) [2]ARC Lab, Tencent PCG [3]Tencent AI Lab [4]CUHK [5]HKUST

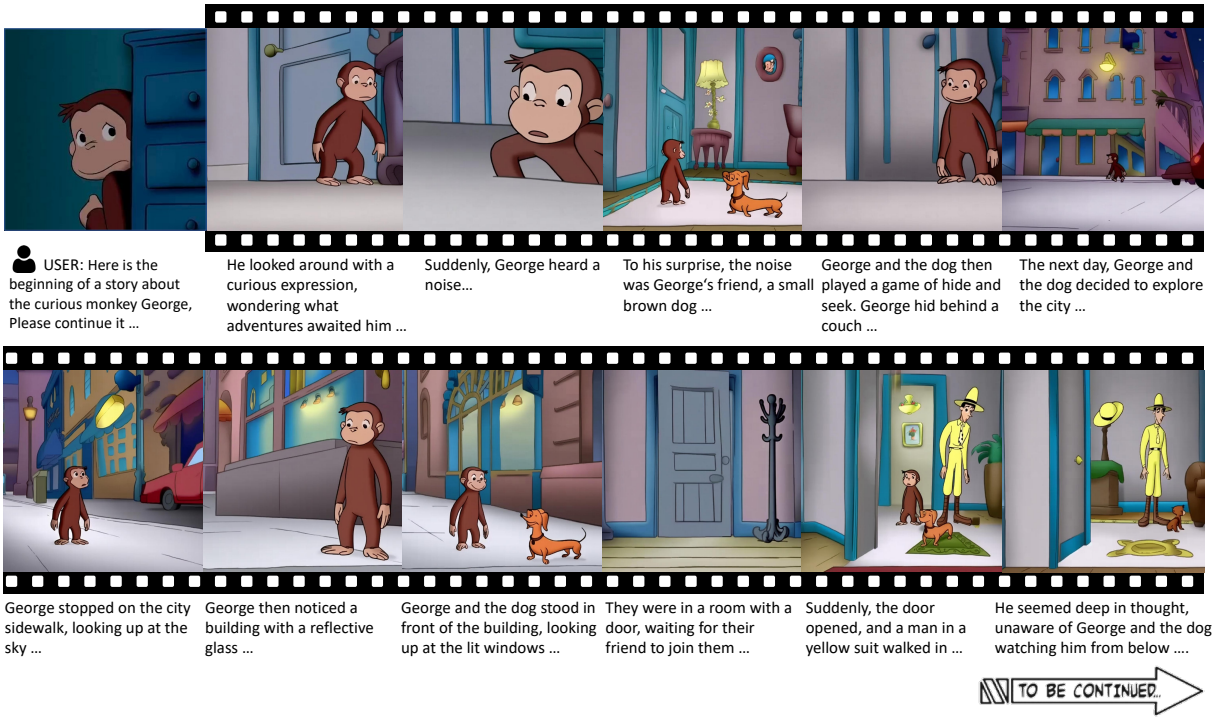

Figure 1. SEED-Story, powered by the MLLM and the multimodal attention sink mechanism, can create long, multimodal stories from a user's starting image and text. It generates coherent narrative text alongside images that consistently represent the characters and style throughout the story. Although the model is trained using up to 10 sequences, it is capable of producing stories that span as many as 25 multimodal sequences (see Appendix).

Abstract

Advances in image generation and open-form text generation have paved the way for tackling the challenging task of multimodal long story generation. In our work, we introduce SEED-Story, a novel approach that extends Multimodal Large Language Models (MLLMs) to generate coherent, extended narratives composed of both interleaved text and images. By leveraging robust MLLMs, our model predicts text tokens and regresses visual features that are subsequently refined through an adapted de-tokenizer, ensuring that generated images consistently depict recurring characters and maintain a unified visual style. Furthermore, we introduce a multimodal attention sink mechanism to overcome the train-short test-long challenge. This mechanism retains recent tokens while preserving critical tokens from both the start and end of image sequences, enabling efficient autoregressive generation of long stories that can extend to 25 sequences, even though training is performed on only 10 sequences. To support our research, we also introduce StoryStream, a large-scale, high-resolution dataset tailored for multimodal long story generation. StoryStream

offers longer narrative sequences and richer visual details than previous datasets, providing a robust benchmark for evaluating image style consistency, story engagement, and image-text coherence. Experimental results demonstrate that SEED-Story produces rich narrative plots and diverse visual scenarios across extended multimodal sequences.

1. Introduction

Interleaved image-text data is ubiquitous on the Internet, characterized by multiple images interspersed with text. In recent years, there has been a surge of interest in generating interleaved image-text content [1, 10, 15, 46, 50], driven by the remarkable advances in image generation [8, 25, 35, 41, 52] and open-form text generation [45, 51, 56].

This has motivated us to tackle a more challenging task: Multimodal Long Story Generation. In this task, the model is given an initial story consisting of narrative text and images, and it must generate extended sequences of text and images to continue the narrative. The goal is to produce a coherent and long narrative where both the text and visuals consistently reflect recurring characters and styles, much like extending a comic book. Achieving this is particularly demanding, as the model must not only understand the intricate interplay between text and images, but also preserve context over a long sequence of story segments, ensuring both semantic depth and visual consistency throughout.

Recently, Multimodal Large Language Models (MLLMs) [6, 21, 22, 25, 38, 55, 57] have demonstrated exceptional ability in understanding multimodal data, making them well-suited for handling interleaved image-text content found in multimodal stories. Building on this strength, we introduce SEED-Story (see Figure 1), a novel approach that not only leverages the robust comprehension capabilities of MLLMs but also enhances them with the ability to generate coherent and extended story sequences that align with the accompanying narrative texts.

Specifically, following previous work [15, 44], we utilize a pre-trained image tokenizer and de-tokenizer, which can decode realistic images with SD-XL [39] by taking the features of a pre-trained ViT as input. During training, we adopt the next-word prediction and image feature regression training objectives to achieve multimodal generation. A fixed number of learnable queries are fed into the MLLM, where the output hidden states are trained to reconstruct the ViT features of the target images. To further enhance the consistency of characters and styles in generated images, we propose de-tokenizer adaptation, where the regressed image features from the MLLM are fed into the de-tokenizer for tuning SD-XL. This adaptation allows for better maintenance of coherence in low-level image details, ensuring a more visually consistent storytelling output.

Furthermore, to efficiently generate long stories, we introduce a multimodal attention sink mechanism. This technique maintains a fixed-size sliding window on the Key-Value cache for the most recent tokens, while also preserving key tokens from the start of both text and image sequences, as well as the end of image tokens. This design effectively overcomes the limitations of window attention when the generated sequence length exceeds the training length, enabling our model to be trained on shorter sequences while generalizing to longer ones. Empirically, our model equipped with the multimodal attention sink mechanism can generate stories with up to 25 multimodal sequences (despite training on only 10 sequences), featuring rich narrative plots and diverse visual scenarios.

Additionally, we introduce a dataset named StoryStream for training and evaluating multimodal story generation. We design an automatic pipeline that leverages MLLMs to obtain a large-scale and high-resolution dataset featuring a sequence of narrative-rich texts and intriguing images, derived from animated videos. StoryStream is four times larger in terms of data volume compared to the existing largest story dataset [26], and it boasts higher image resolution, longer sequence lengths, and more detailed story narratives. We further meticulously design evaluation metrics to assess multimodal story generation, taking into account image style consistency, story text engagement, and image-text coherence. The evaluation results demonstrate that our model, SEED-Story, achieves superior performance in these aspects.

In summary, Our contributions are three-fold. (1) We propose SEED-Story, a novel method that leverages an MLLM to generate multimodal long stories with rich narrative text and contextually relevant images. (2) We propose multimodal attention sink to enable the efficient generation of long stories with sequence lengths larger than those used during training. (3) We introduce StoryStream, a large-scale dataset specifically designed for training and benchmarking multimodal story generation.

2. Related Work

Personalized Story Visualization v.s. Multimodal Long Story Generation In addition to conventional story visualization [16, 23, 30–32, 34, 40], there are two different types of story generation in recent studies. Figure 2 presents the main differences between these two tasks.

Personalized Story Visualization, as explored in works such as [5, 26, 29, 42, 47], focuses on generating images that depict specific characters performing various actions or appearing in different scenes, all based on the provided captions. This approach follows the conventional text-to-image generation paradigm with the following key characteristics: 1) The generation process uses the text captions directly. 2) Each image is generated without dependency on the preceding images; the process is non-sequential, meaning that the

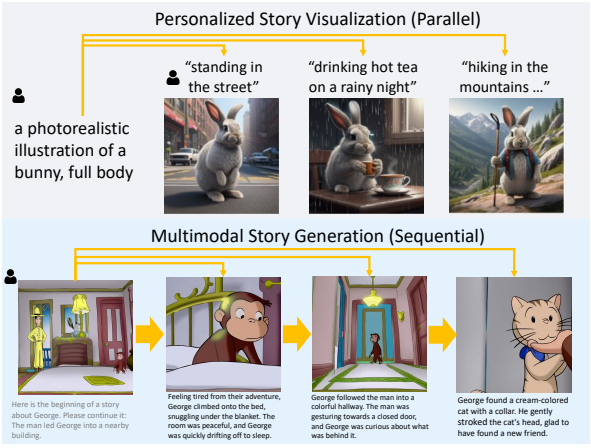

Figure 2. Comparison between personalized story visualization and multimodal story generation. **Upper**: a bunch of captions are given for consistent text-to-image generation. All images are generated in paraleel, depending merely on the user's input. **Lower**: Our model sequentially creating a consecutive storylines along with their corresponding images. Outputs are based on all the previous images and texts.

narrative is not constructed as a continuous series. 3) While these methods stress the ability to generalize to unseen characters, this comes at the cost of flexibility. Every generated image must prominently feature the main character, limiting the dynamic representation of diverse scenes.

In contrast, Multimodal Story Generation, as discussed in [43] and ours, seeks to produce a sequence of consecutive storylines paired with their corresponding images, much like a serialized comic. This task requires the model to predict coherent story developments and generate matching illustrations by integrating previously generated outputs as context in an auto-regressive manner. The salient features of this approach include: 1) The model itself is responsible for generating the narrative text. 2) The story is built in a sequential manner, with each segment and its illustration relying on the context provided by earlier parts of the story. 3) This approach offers greater flexibility in accommodating diverse scenes and narrative turns. It allows for variations where not every image must include a clear depiction of the main character. As a result, it inevitably tends to exhibit in-domain generalization. Figure G of the appendix demonstrates this ability. Given the same initial image, our model produces different storylines based on varying initial narratives. However, these approaches are still limited to seen characters.

The task of multimodal story generation presents a more substantial challenge, and we follow previous research [43] to adopt a closed-set setting. We believe the ultimate goal of multimodal story generation should be to generate highly diverse scenarios while also generalizing to unseen characters, which will be explored in the future work.

MLLM for Multimodal Story Generation In the rapidly evolving domain of large language models (LLMs) [7, 9, 51] and multimodal large language models (MLLMs) [6, 10, 13, 14, 20, 22, 24, 25, 27, 28, 38, 44, 53, 55, 57, 58], recent work StoryGPTV [43] explores using MLLMs for story generation by converting visual features into token embeddings for image generation, but requires aditional character and object masks for training. MM-interleaved [50] designs a multi-scale and multi-image feature synchronizer module (MMFS) to process interleaved text-image data and generates next image conditioned on the previous context features from LLM. But this MMFS module makes it difficult to generate long multimodal stories due to the complex multi-scale attention mechanism. In this work, we design a novel model SEED-Story to generate long multimodal stories without extra mask guidance.

3. Method

In this section, we first introduce the training pipeline of our multimodal story generation model. Then, we propose multimodal attention sink to enable long story inference.

3.1. Multimodal Story Generation Training

Visual Tokenization and De-tokenization The overview of our training pipeline is presented in Figure 3. To effectively extend visual stories, our model must comprehend and generate both images and text. Drawing inspiration from recent advancements in generative MLLMs that unify image comprehension and generation [39], we develop a multimodal story generation model. Our model employs a pre-trained Vision Transformer [11] (ViT) as the visual tokenizer and a pre-trained diffusion model as the visual de-tokenizer to decode images by using ViT's features as inputs. Specifically, visual embeddings from the ViT tokenizer are fed into a learnable module, which then serves as inputs for the U-Net of the SD-XL [39]. This process replaces the original text features with visual embeddings. During this stage, the parameters are optimized using open-world text-image pair data as well as story data to enhance the model's encoding-decoding capability. After this training phase, we expect the visual tokenizer and de-tokenizer modules to preserve as much image information as possible in the feature space.

Story Instruction Tuning In our instruction tuning process for story generation, we sample a random-length subset of a story data point for each iteration. The model is tasked with predicting the next image and the next sentence of the story text. Within MLLM, all images are converted

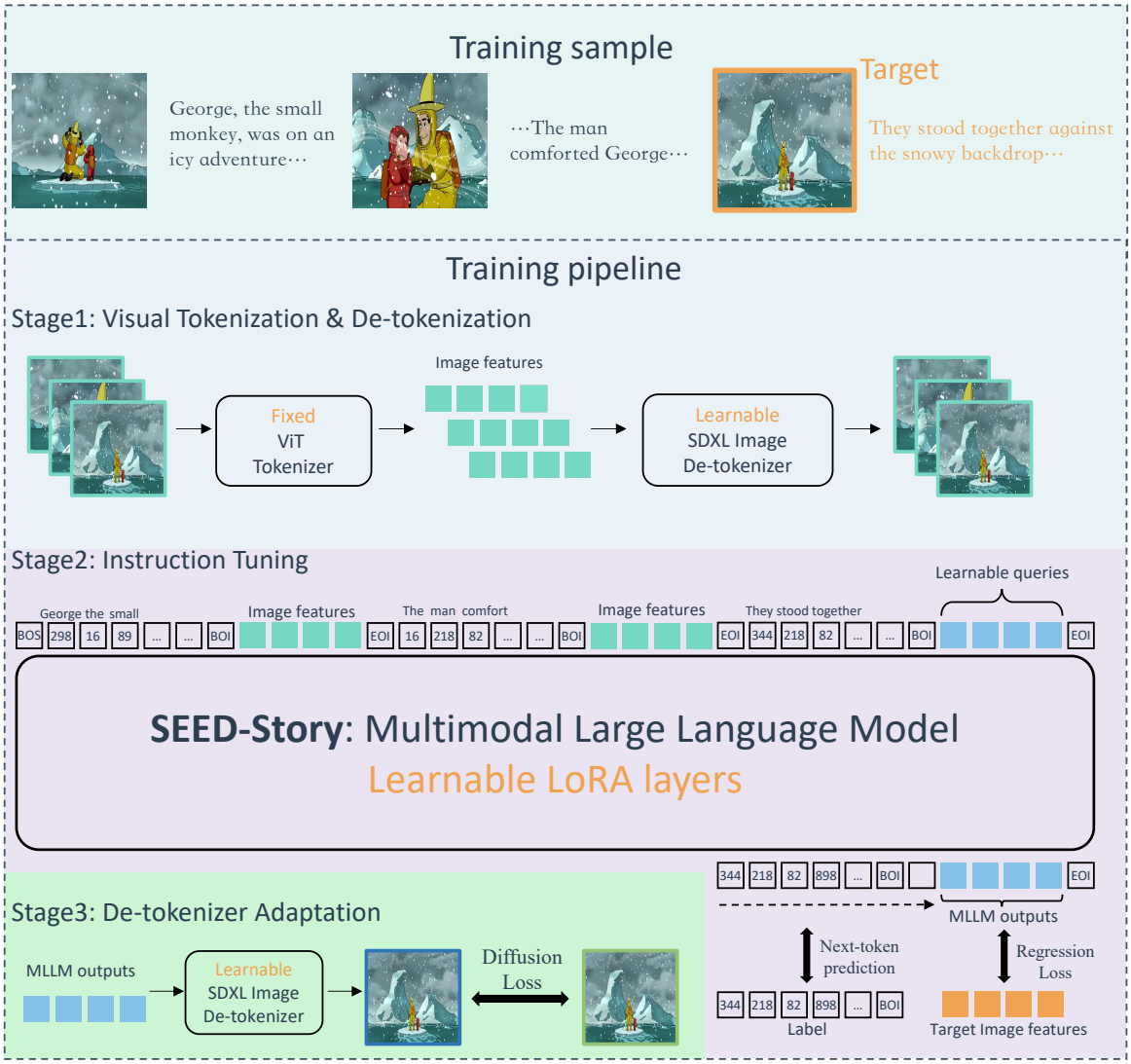

Figure 3. Overview of the SEED-Story Training Pipeline: In **Stage 1**, we pre-trains an SD-XL-based de-tokenizer to reconstruct images by taking the features of a pre-trained ViT as inputs. In **Stage 2**, we sample an interleaved image-text sequence of a random length and train the MLLM by performing next-word prediction and image feature regression between the output hidden states of the learnable queries and ViT features of the target image. In **Stage 3**, the regressed image features from the MLLM are fed into the de-tokenizer for tuning SD-XL, enhancing the consistency of the characters and styles in the generated images.

into image features using a pre-trained ViT tokenizer. For the target text tokens, we perform next-token prediction and use Cross Entropy loss to train for this discrete target. For target image features, the model uses a series of learnable queries as inputs and continuously outputs a series of latents. We then compute the cosine similarity loss between the MLLM's output and the target image features. During this stage, we fine-tune the SEED-Story model using a LoRA Hu et al. [18] module.

De-tokenizer Adaptation After instruction tuning, the SEED-Story MLLM effectively produces story images with correct semantics but lacks style consistency and details. We attribute this issue to the misalignment between the latent space of the MLLM output and the image features. To address this, we perform de-tokenizer adaptation for style and texture alignment. In this stage, only the SD-XL image de-tokenizer is trained. Conditioned on the MLLM output embeddings, SD-XL is expected to generate images

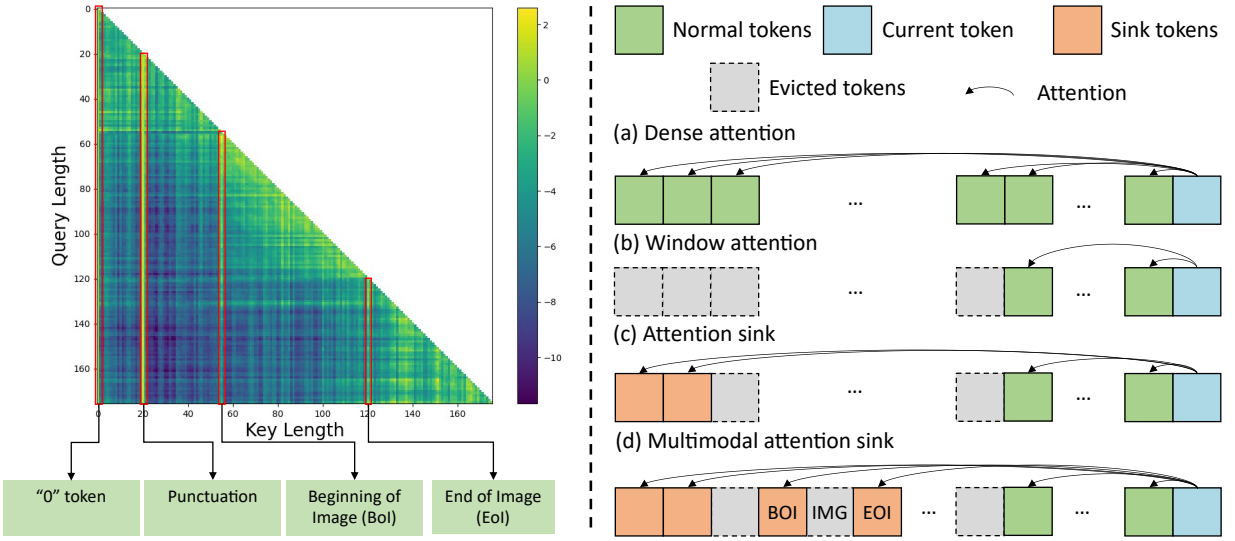

Figure 4. **Left:** Visualization of the attention map when predicting the next token for multimodal story generation. We observe that important attentions are aggregated into the first token of the whole sequence ("0" token), punctuation tokens, tokens adjacent to BoI, and tokens adjacent to EoI. **Right:** The diagram of (a) Dense attention, which preserves all tokens in KV cache. (b) Window attention, which evicts preceding tokens by a sliding window. (c) Attention sink, which preserves the beginning tokens based on window attention. (d) Multimodal attention sink, which preserves the beginning of text tokens, images tokens, and the end of image token based on window attention. It can efficiently enable our model to generalize to generating longer sequences than the training sequence length.

that are pixel-level aligned with the ground truth. The separate training of the de-tokenizer offers two key advantages. First, it avoids optimization conflicts between the LLM and the de-tokenizer. Second, it conserves memory, making the process executable on GPUs with limited memory. Please see more analysis in Section 5.3 of our appendix.

Table 1. Key Token Occurrence for Various Tokens

BOS	**IMG57**	**EOI**	**IMG04**
4320	4320	4320	4140
","	**IMG60**	**IMG61**	**IMG59**
4140	4120	3730	3132
IMG62	**IMG63**	**BOI**	**IMG56**
1651	607	603	361

3.2. Long Story Generation

Generating long visual stories has substantial potential in various applications, including education and entertainment. However, creating these stories with MLLMs presents significant challenges. Datasets for extended, interleaved stories are not only rare but also impede the training process due to their complexity. To address this, we have to employ a train-short-test-long approach, training models on shorter narratives and extending to longer generations during inference.

Moreover, during inference, generating significantly longer stories than the training data often leads to model degradation, producing lower-quality images, as illustrated in the first row of Figure 6. This process also requires extensive token usage to ensure continuity and coherence, which in turn increases memory and computational demands.

A simplistic solution for this is to use a sliding window technique, depicted in Figure 4 right (b). However, this method disrupts the token relationships in the Key-Value cache, resulting in subpar generative outcomes, as demonstrated by StreamingLLM [54]. To overcome this, StreamingLLM introduces an attention sink mechanism that preserves the initial tokens, thus allowing for efficient processing of lengthy generations without quality compromise. While effective in language models, its efficacy diminishes in multimodal contexts, as shown in Figure 4 right (c).

To address this limitation in multimodal contexts, we revisit the attention maps of MLLMs. We analyze 5600 attention maps from various models, layers, and input sequences, to identify "key tokens" with the highest mean attention values. For each attention map, we computed the mean attention value across the query dimension for every key and selected the top 10 keys. We then aggregated these results to determine how often each token appeared as a key token. Table 1 shows the tokens with the highest occurrences, with most queries focusing on four key token categories: 1)

Table 2. Comparison of multimodal story generation datasets. The table provides details on the number of images(# Images), their resolution, the total length of visual stories, and the average text length per sentence, which indicates the narrative detail of the text. Note that StorySalon has various resolution of images, and we choose one of the typical sizes presented here.

Datasets	#Images	Resolution	Story Len.	Text Len.
Flintstones [17]	122,560	128×128	5	86
Pororo [23]	73,665	128×128	5	74
StorySalon [26]	159,778	432×803	14	106
StoryStream	257,850	480×854	30	146

Starting tokens (BOS). 2) Punctuation (“,” “.” and “;” ...). 3) Image tokens near BOI (BOI, IMG04). 4)Image tokens near EOI (from IMG57 to IMG63, added EOI). Our analysis reveals that unlike language-only models, MLLMs place considerable attention on specific image tokens, particularly those near the BoI and EoI, as illustrated in Figure 4 left.

Building on these insights, we propose a new mechanism for extended generation in MLLMs, termed the multimodal attention sink. During generation, we consistently retain the starting tokens and the image tokens adjacent to the BoI and EoI. Although punctuation tokens receive high attention values, their latent value norms are minimal, contributing insignificantly to the final output, so we do not keep them, as noted by [12]. Our proposed mechanism enables our model to generate high-quality images while maintaining a low computational footprint.

4. StoryStream Dataset

4.1. Dataset Construction

An ideal source for creating a multimodal story generation dataset is cartoon series, which inherently contain rich plots and consistent character portrayals. We selected three cartoon series to construct our dataset and we present the Curious George in the main body of our paper. The process begins with collecting various series, from which we extract keyframes and their associated subtitles [19]. Each keyframe is then processed by GPT-4V [33] or Qwen-VL [6] to generate a detailed image description. These elements—keyframe, subtitle, and description—are compiled into a single group. We aggregate 30 such groups and input them into GPT-4, supplemented with background information about the cartoon series. Following our instructions, GPT-4 generates high-quality narrative texts suitable for training story generation models.

During dataset construction, we discovered that employing the above chain of thought approach not only produces more accurate narrative text but also speeds up the construction process. Unlike directly feeding all images directly to GPT-4, which is limited to 10 images due to API constrains, our approach produces longer stories. We also significantly improve the model’s understanding of each image by incorporating detailed descriptions. This enhancement in image comprehension enriches the narrative details, providing a richer story generation reference.

4.2. Key Features

Large-scale. Our StoryStream dataset comprises three subsets totaling 257,850 images. This represents a significant improvement over existing datasets in terms of scale, specific numbers are presented in Table 2. To the best of my knowledge, it is the largest visual story generation dataset featuring consistent main characters.

High Resolution. Unlike existing story generation datasets which offer images at a resolution of 128x128, our story stream dataset provides high-resolution images of 480x768.

Narrative Text. Our dataset diverges from existing ones that utilize simple and descriptive language. We offer abstract, narrative, detailed, and story-toned texts that are more akin to real-world applications, such as visualizing narratives from a storybook, examples are shown in Figure B of the Appendix. Story text of existing datasets obey the form of “name” + “action”, like “Poby is playing the violin.”. Contrarily, our story text involves more intrinsic elements. This effectively enhances the engagement of audiences. An analysis of the average text length per sentence is shown in column 5 of Table 2.

Long Sequence. Moreover, our dataset enhances long story comprehension by offering up to 30 images per story point. Within these 30 images, our corresponding texts present a cohesive narrative, effectively conveying the progression and intricacies of extended stories.

5. Experiment

Table 3. The FID evaluation between SD-XL, StoryGen, MM-Interleaved, and SEED-Story.

Model	SD-XL	StoryGen	MM.	SEED.
FID ↓	67.29	73.42	65.71	**62.23**

5.1. Multimodal Story Generation

Comparison For the quantitative evaluation of multimodal story generation, we first fine-tune the recently developed MM-Interleaved model on our training dataset as a baseline. We employ three categories of metrics to evaluate these models. Fréchet Inception Distance (FID) is used as a quantitative measure to assess the visual quality of generated images, as shown in Table 3. Additionally,

Table 4. The GPT-4V evaluation results between SD-XL, StoryGen, MM-Interleaved, and SEED-Story. Consi. denotes consistency. Txt Engage denotes the level of engagement of story text. Coher. denotes coherence. MM. is MM-interleaved. SEED. is our SEED-story.

Metric	SD-XL	StoryGen	MM.	SEED.
Style Consi. ↑	7.12	**8.66**	7.42	8.61
Text Engage ↑	-	-	6.27	**6.31**
Coher. ↑	8.11	5.21	6.45	**8.24**

Table 5. The human evaluation results between SD-XL, StoryGen, MM-Interleaved, and SEED-Story. Img. Qual. denotes image quality. Consi. denotes consistency. Div. denotes diversity. Txt Engage denotes the level of engagement of story text. Coher. denotes coherence. MM. is MM-interleaved. SEED. is our SEED-story.

Metric	SD-XL	StoryGen	MM.	SEED.
Img. Qual. ↑	**8.90**	6.34	8.75	8.68
Style Consi. ↑	5.25	8.52	**8.92**	8.65
Img. Div. ↑	**9.00**	6.15	7.91	8.42
Text Engage ↑	-	-	5.43	**6.26**
Coher. ↑	7.50	7.75	7.31	**8.30**
Preference ↑	3.5%	13.5%	29.5%	**53.5%**

we leverage GPT-4V ("gpt-4-turbo-2024-04-09") to score the generation results across several dimensions: (a) Image Style Consistency – evaluates the uniformity of visual style across different images. (b) Story Text Engagement – measures the ability of the narrative to captivate and sustain audience interest. (c) Image-Text Coherence – assesses the alignment and relevance between images and their corresponding texts. The result is shown in Table 4. Details of this GPT-4V evaluation are provided in Section F of the Appendix. Furthermore, we conduct a user study where participants rate the models across five aspects: image quality, image style consistency, image diversity, text engagement, and image-text coherence. We finally ask the participants to choose their preferrence based on the above scores. The results in Figure 5 were obtained from 80 participants. All of the quantitative evaluations above demonstrate that SEED-Story either outperforms or shows competitive results to the baseline models, and it particularly presents a clear advantage in cross-domain image-text coherence. Figure K and L in Section E shows more visual results. For qualitative results, we present them in the Appendix. Please see Section D.

Generalization SEED-story can generate novel, extended narratives and illustrations from user prompts, showcasing effective generalization for in-domain data. Users can control the plot with text, leading to unique stories within a learned world, such as its "George" character set, as shown in Figure J. Its primary limitation is the inability to generate characters from outside its training domain, like those from other animations.

Figure 5. Generated story images with and without the 3rd stage: **de-tokenizer adaptation**. We shorten it as D.A. in this figure. Although the generated images without de-tokenizer adaptation preserve semantic information, they display low-quality textures and inconsistency in cartoon style. As shown in the images, the monkey's face changes across different images, and the quality of details remains low. Our de-tokenizer adaptation effectively enhances these aspects.

5.2. Story Visualization

Previous story generation approaches primarily utilize diffusion models, focusing on visualizing story images. These models take the previous image and text as input, and then generate only the next image based on the current text prompt. For a fair comparison, we adapt our model to a visualization-only format. For StoryGen [26], we also train it to produce images with previous images and texts. For LDM [41], we only give it text-image pairs. We conduct a quantitative evaluation in Table 6 to demonstrate our effectiveness. SEED-Story model shows better style and character consistency and higher quality compared to baselines. More story visualization results are shown in Section D of our appendix.

Table 6. Quantitative evaluation for story visualization.

Model	**FID ↓**	**CLIP Score↑**
LDM	67.29	0.7585
StoryGen	73.74	0.7573
SEED-Story	**67.01**	**0.7793**

5.3. Ablation on De-tokenizer Adaptation

We validate the effectiveness of de-tokenizer adaptation here. As shown in Figure 5, generated images without the de-tokenizer adaptation stage exhibit semantic relevance with consistent backgrounds and characters, thanks to MLLM's context preservation. However, they suffer from texture distortion and inconsistency in style. With

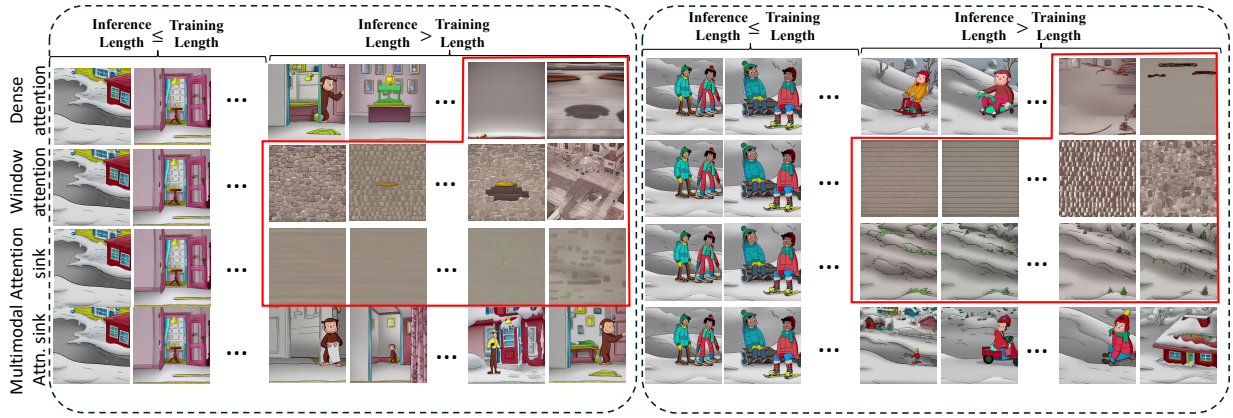

Figure 6. The visualization of generating long stories with different attention mechanisms. Without multimodal attention sink, MLLM cannot generate meaningful long image sequences. As highlighted in the red boxes, other methods produce meaningless images in the later frames.

Table 7. FID scores with and without the de-tokenizer adaptation stage. We shorten de-tokenizer adaptation as D.A. in this table.

Model	FID
w/o D.A.	153.93
w/ D.A.	99.79

de-tokenizer adaptation, the images show improved consistency in style and character appearance. The calculated FID scores in table 7 confirm that de-tokenizer adaptation enhances image quality.

5.4. Ablation on Multimodal Attention Sink

To verify the effectiveness of multimodal attention sink in long story generation, we conduct an experiment visualizing a long story using the SEED-Story model, but with varying attention mechanisms. We chunk our data into stories of length of 10 considering the training efficiency. We set the window size as the same as the training length. Qualitative results presented in Figure 6 demonstrate that window attention quickly collapses when the inference length exceeds the training length. We provide more cases in our appendix. Both dense attention and attention sink approaches fare better, yet still fail to produce meaningful images as the inference sequence lengthens. In contrast, the multimodal attention sink consistently produces high-quality images. In terms of efficiency, the multimodal attention sink exhibits significant improvement over dense attention, with only a modest increase in time and memory costs compared to window attention and vanilla attention sink. These additional costs stem from retaining extra image tokens in the KV cache. Quantitative results presented in Table 8 substantiate the above conclusion.

Table 8. Quantitative evaluation of long story generation with various attention mechanisms. FID and CLIP scores are calculated by comparing ground truth images with generated images. Inference time and memory usage are calculated by generating 50 sequences multiple times for average.

Metrics	FID ↓	CLIP Score ↑	Inference Time (s) ↓	Memory (GB) ↓
Dense Attn	119.72	0.705	569.67	37.99
Window Attn	334.90	0.598	**450.64**	**30.81**
Attn Sink	221.53	0.676	451.94	**30.81**
Ours	**79.67**	**0.728**	473.98	31.82

6. Conclusion

This work introduces SEED-Story, a pioneering approach that leverages a Multimodal Large Language Model (MLLM) to generate cohesive, multimodal long stories. To achieve this, we propose an innovative multimodal attention sink mechanism, enabling the model to efficiently generalize to generating long sequences far exceeding the length of its training data. Furthermore, we present StoryStream, a new, high-quality dataset specifically designed for training and benchmarking multimodal story generation. Derived from children's cartoons, StoryStream provides the lengthy, coherent image and text sequences crucial for developing robust story generation models. Our comprehensive evaluations demonstrate that SEED-Story significantly advances the state-of-the-art in creating engaging and continuous multimodal narratives.

References

[1] Emanuele Aiello, Lili Yu, Yixin Nie, Armen Aghajanyan, and Barlas Oguz. Jointly training large autoregressive multimodal models. *arXiv preprint arXiv:2309.15564*, 2023. 2

[2] Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models, 2023. 13

[3] Animaj. Animaj official website, 2024. Accessed: 2024-05-22. 16

[4] Animaj. Rabbids invasion official youtube channel, 2024. Accessed: 2024-05-22. 16

[5] Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. The chosen one: Consistent characters in text-to-image diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 2

[6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 2, 3, 6

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901, 2020. 3

[8] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 2

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 3

[10] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023. 2, 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[12] Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive kv cache compression for llms. *arXiv preprint arXiv:2310.01801*, 2023. 6

[13] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023. 3

[14] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer, 2023. 3

[15] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation, 2024. 2

[16] Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. Talecrafter: Interactive story visualization with multiple characters. *arXiv preprint arXiv:2305.18247*, 2023. 2

[17] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 598–613, 2018. 6

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4

[19] Maciej Kilian, Romain Beaumont, Daniel Mendelevitch, Sumith Kulal, and Andreas Blattmann. video2dataset: Easily turn large sets of video urls to a video dataset. https://github.com/iejMac/video2dataset, 2023. 6

[20] Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *NeurIPS*, 2023. 3

[21] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023. 2, 3

[23] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6329–6338, 2019. 2, 6

[24] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 3

[25] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 2, 3

[26] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, and Weidi Xie. Intelligent grimm–open-ended visual storytelling via latent diffusion models. *arXiv preprint arXiv:2306.00973*, 2023. 2, 6, 7

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 3

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 3

[29] Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fhad Khan, Shiqi Yang, Yaxing Wang, Jian Yang, and Mingming Cheng. One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[30] Adyasha Maharana and Mohit Bansal. Integrating visuospatial, linguistic and commonsense structure into story visualization. *arXiv preprint arXiv:2110.10834*, 2021. 2

[31] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Improving generation and evaluation of visual stories via semantic consistency. *arXiv preprint arXiv:2105.10026*, 2021.

[32] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European Conference on Computer Vision*, pages 70–87. Springer, 2022. 2

[33] OpenAI. Gpt-4v: Optimizing language models for dialogue. https://www.openai.com/chatgpt, 2023. 6

[34] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2920–2930, 2024. 2

[35] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, 2021. 2

[36] PBS Kids. Curious george official website, 2024. Accessed: 2024-05-22. 15

[37] PBS Kids. Curious george official youtube channel, 2024. Accessed: 2024-05-22. 15

[38] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2, 3

[39] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 3

[40] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2493–2502, 2023. 2

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 7

[42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2

[43] Xiaoqian Shen and Mohamed Elhoseiny. Storygpt-v: Large language models as consistent story visualizers, 2023. 3

[44] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. 2023. 2, 3

[45] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 2

[46] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv e-prints*, pages arXiv–2405, 2024. 2

[47] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024. 2

[48] TheLandBeforeTime. The land before time official youtube channel, 2024. Accessed: 2024-05-22. 16

[49] TheLandBeforeTime. The land before time official website, 2024. Accessed: 2024-05-22. 16

[50] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024. 2, 3

[51] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 3

[52] Luozhou Wang, Shuai Yang, Shu Liu, and Ying cong Chen. Not all steps are created equal: Selective diffusion distillation for image manipulation, 2023. 2

[53] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 3

[54] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 5

[55] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 2, 3

[56] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 2

[57] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2, 3

[58] Jinguo Zhu, Xiaohan Ding, Yixiao Ge, Yuying Ge, Sijie Zhao, Hengshuang Zhao, Xiaohua Wang, and Ying Shan. Vl-gpt: A generative pre-trained transformer for vision and language understanding and generation. *arXiv preprint arXiv:2312.09251*, 2023. 3