Appendix

A. Implementation Details

A.1. Visual Tokenization and De-tokenization

For visual tokenization, we use Qwen-VL pre-trained ViT-G. We first resize the image to 448x448 images and then use ViT to produce its feature of length256 with 4096 dimension. (shape: [256, 4096]). Inside the MLLM, we use a Q-Former architecture to process the image embedding. It takse the ViT image feature as key and Value, and conduct attention with its learnable queries. The length of learnable queries are 64. For de-tokenization, we also use a Q-Former architecture to transform the MLLM output to the shape of SD-XL condition embedding.

A.2. Instruction Tuning

Instruction tuning data is formatted as follows: for each story, we sample a random length and compute losses on the last sequence (highlighted in red text). The sequence format is structured as:

```
<bos>[start of the story.][User
prompt:  ][following sequence
1][following sequence
2][following sequence
3][following sequence 4] ...
[target sequence]<eos>
```

For our language model (LLM), we utilize the LLAMA2-7B pre-trained model and finetune it using LoRA, supported by the *peft* library. The hyper-parameter r is set to 6, and lora_alpha is set to 32. The modules optimized include the q_projection_layer, v_projection_layer, k_projection_layer, o_projection_layer, gate_projection_layer, down_projection_layer, and up_projection_layer. We employ a learning rate of 1×10^{-4} to finetune this model on our dataset across approximately 6 epochs, utilizing 8 NVIDIA-A800 GPUs.

A.3. De-tokenizer Adaptation

In this stage we fully finetune the SD-XL model. The data format is as the same as instruction tuning, but we fix all MLLM params and optimize only the SD-XL. It takes the MLLM output and is asked to produce image correspond to the ground truth. The SD-XL model was trained using 4 NVIDIA-A800 GPUs. A learning rate of 1×10^{-4} was chosen to facilitate gradual weight updates, ensuring stable convergence, while a weight decay of 0.03 was applied for regularization to prevent overfitting. Training was performed using mixed precision (`bf16`), which significantly reduced memory usage and accelerated the training process without compromising the model's accuracy. The model underwent three training epochs, balancing the learning of complex patterns against computational resource use, optimized for large-scale datasets and sophisticated model architectures.

B. Analysis of Multimodal Attention Sink

B.1. Attention Map Visualization

In this section, we present additional visualizations of attention maps. These maps are derived from various model runs, including varying data lengths, attention heads, and layers. The visualizations consistently reveal a pattern of attention focused on "0" tokens, punctuation, tokens adjacent to Begin-of-Image (BoI), and tokens adjacent to End-of-Image (EoI).

C. StoryStream Dataset Visualization

Figure B compares our StoryStream dataset with baseline datasets (FlintstonesSV and PororoSV). Our dataset surpasses existing ones in scale, resolution, and sequence length. In terms of textual content, our dataset deviates from prior works that primarily rely on simple, descriptive language. Instead, we provide abstract, narrative-driven, and richly detailed story texts that closely resemble real-world storytelling. Existing datasets typically follow a rigid "character + action" format, such as "Poby is playing the violin." In contrast, our dataset incorporates deeper narrative elements, enhancing the expressiveness and complexity of story generation.

To efficiently generate storylines, we employ a combination of multimodal models (GPT-4V and Qwen-VL) alongside pure language models. This hybrid approach ensures both contextual coherence and high-quality storytelling.

D. Qualitative Comparisons

Table C and Table D demonstrate our effectiveness when compared to story visualization method, including LDM and StoryGen.

More visual results are shown in Figure D. SEED-Story model shows better style and character consistency and higher quality compared to baselines. We also showcase the visualization result of our model on Rabbids Invasion and The Land Before Time. Please see Figure H and I.

Table E and Table F demonstrate our effectiveness when compared to multimodal story generation baseline, MM-Interleaved.

E. Multimodal Story Generation Results

In this section, we present more multimodal story generation results of our SEED-Story. It keeps produce story image and text with high quality. We demonstrate that our SEED-story can effectively generate stories with different

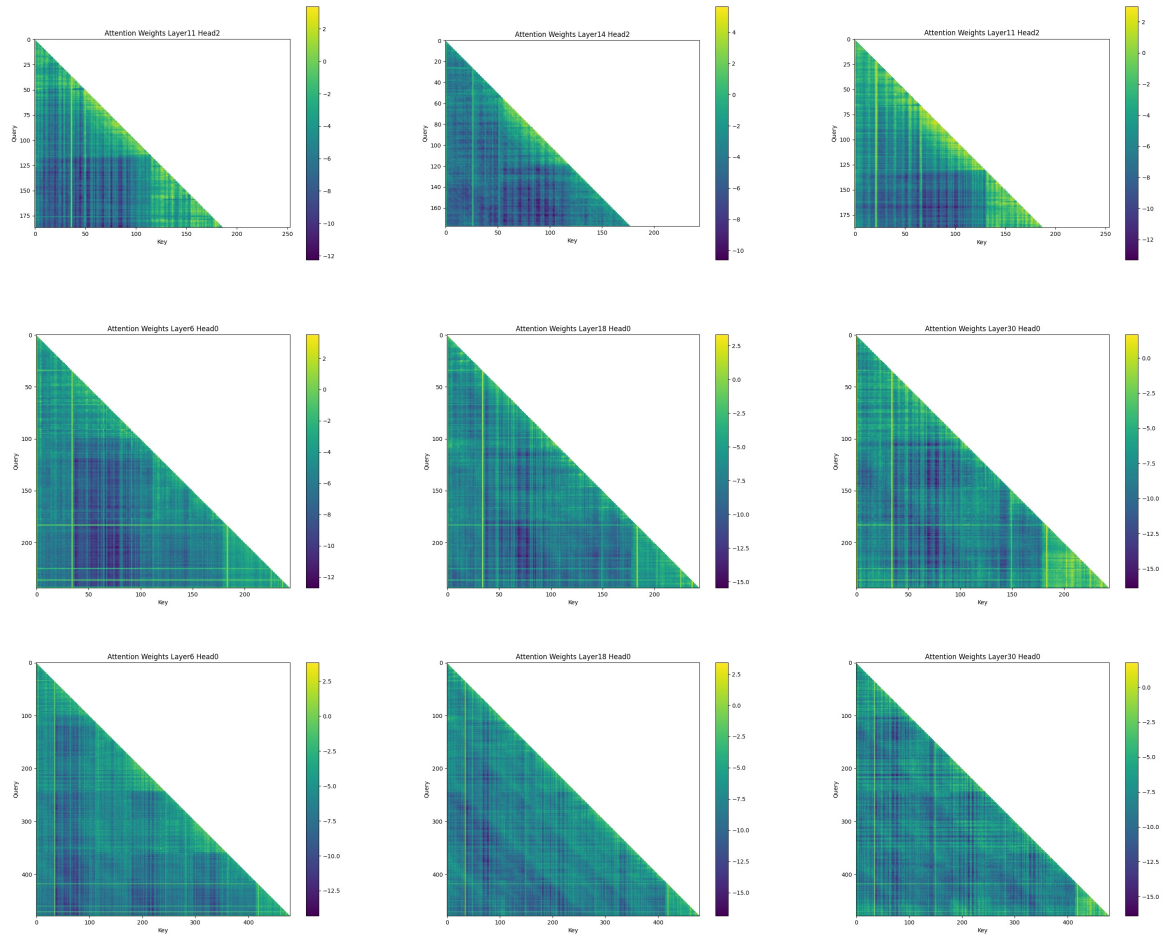

Figure A. Visualization of attention maps from various model runs, showcasing attention patterns across different data lengths, attention heads, and model layers. Notably, the maps highlight consistent focus on '0' tokens, punctuation, tokens adjacent to Begin-of-Image (BoI), and tokens adjacent to End-of-Image (EoI).

plots and corresponding illustrations based on the different377 beginnings provided by the user, as shown in Figure J. Figure K and Figure L prove our multimodal long story generation capabilities. SEED-story can generate long sequences with engaging plots and vivid images.

F. Details about GPT-4V Evaluation

F.1. Comparative Evaluation

To evaluate the effectiveness of MM-interleaved and SEED-Story in multimodal story generation, we initiate an experiment where each model produces a story of five segments, based on a common starting image and text. The segment limit is set to five to accommodate the constraints of GPT-4V, which can handle a maximum of ten images per input session. In total, we generate 180 stories for assessment. For evaluation, we employ GPT-4 or GPT-4V to determine which model produces the better story in each case, based on the framework established in L-Eval [2]. We calculate the win rate for each model to determine its performance relative to its counterpart. The prompt we used is shown below.

> "Please act as an impartial judge and evaluate the quality of the generation story contents provided by two AI assistants. Your job is to evaluate

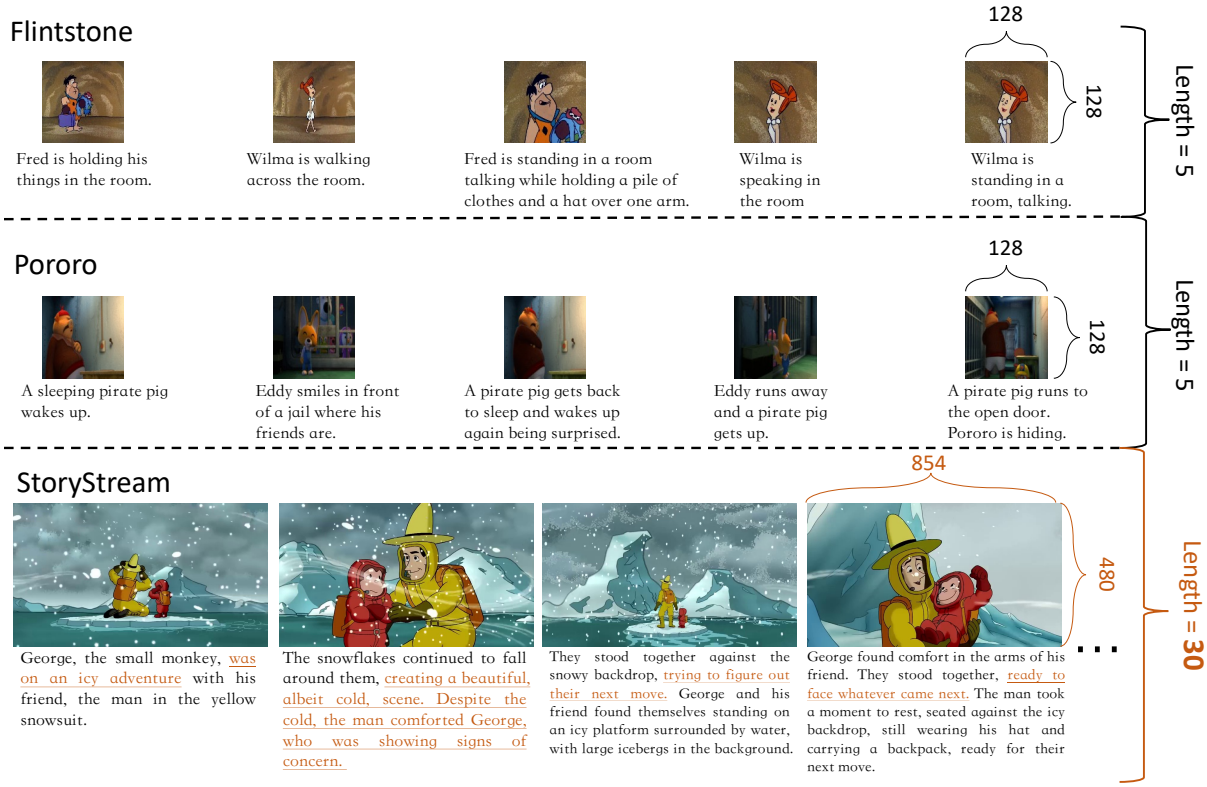

Figure B. Data sample of our StoryStream dataset and existing multimodal story generation datasets. Our multimodal story sequences consist of high-resolution images that are visually engaging, and detailed narrative texts as underlined, closely resembling the real-world storybooks. Additionally, our stories are more extended in length.

which assistant's generation is better. Your evaluation should consider **{the style consistency of the story images / the engagement of the story / the coherence of the generated text and images}**. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants.Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie."

F.2. Score Evaluation

We also provide a prompt for directly estimating the performance of the generated results without comparing to others. The prompt we used is shown below.

"Please act as an impartial judge and evaluate the quality of the generation story contents provided by an AI assistant. Your job is to give a score out of 10. Your evaluation should consider **{the style consistency of the story images / the engagement of the story / the coherence of the generated text and images}**. Do not allow the length of the responses to influence your evaluation. Be as objective as possible. After providing your explanation, output your final score by strictly following this format: "[[score]]", such as "[[7]]"."

G. Story Video

To showcase the capabilities of our multimodal generation model, we employ a video generation technique to animate the images. We then synchronize these moving images with audio to create a narrative video, which is available in our supplementary materials.

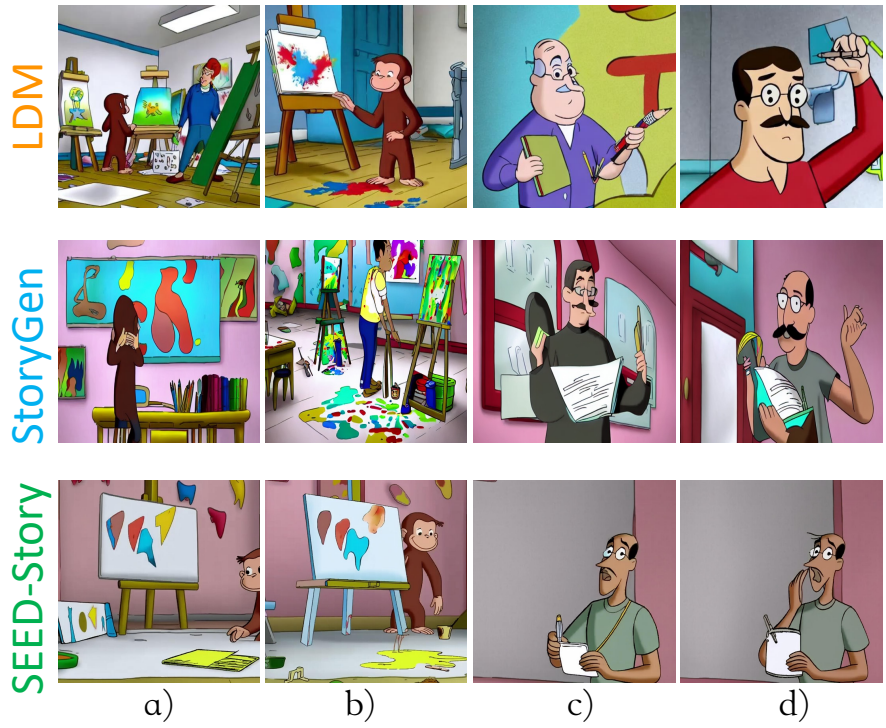

Figure C. Story visualization comparison between SEED-Story and baseline models. SEED-Story generates images with higher quality and better consistency.

H. Data Usage and License

H.1. Curious George

Curious George is an animated series featuring George, a curious monkey whose adventures teach preschoolers about math, science, and engineering. Guided by The Man with the Yellow Hat, George explores the world through problem-solving and experimentation, making it a delightful and educational experience for young viewers.

Curious George is released on PBS KIDS [36, 37], a not-for-profit institution. It is a production of Imagine, WGBH and Universal. Curious George and related characters, created by Margret and H.A. Rey, are copyrighted and trademarked by Houghton Mifflin Harcourt and used under license. Licensed by Universal Studios Licensing LLC. Television Series: ©2024 Universal Studios. The terms of use of them are provided in https://www.pbs.org/about/about-pbs/terms-of-use/.

Our usage fully comply with the terms of use. 1) Personal Uses Permitted: My project is non-commercial and educational, which aligns with personal uses as outlined by PBS. we are not using the information for commercial purposes or exploiting it in a manner inconsistent with PBS rules. The use is strictly for educational and research purposes within an academic setting. 2) User's Obligation to Abide By Applicable Law: We will ensure all research activities comply with local laws, particularly those relating to copyright and intellectual property rights. Our use will not involve unauthorized reproduction, distribution, or exhibition that violates Intellectual Property Laws. All data are for research only. 3) Content of Information: We will responsibly use the "Curious George" materials, ensuring that all content used in our research is accurately cited and acknowledged. Any PBS content incorporated into your project will be clearly attributed to PBS.

H.2. Rabbids Invasion

"Rabbids Invasion" is a French-American computer-animated TV series that breathes life into the zany antics of Ubisoft's popular Rabbids video game characters. Created by Jean-Louis Momus and featuring the voice of Damien Laquet, the show is a dynamic blend of humor and adventure tailored for a family audience. Since its debut on August 3, 2013, on France 3, the series has enjoyed multiple seasons and a global reach. The Rabbids are mischievous rabbit-like creatures whose escapades lead them

into all sorts of unpredictable and hilarious situations, making "Rabbids Invasion" a delight for both kids and adults alike. Thanks to their release, we derive some subsets from the cartoon series Rabbids Invasion [3, 4].

H.3. The Land Before Time

The Land Before Time, an iconic animated film series created by Judy Freudberg and Tony Geiss and distributed by Universal Pictures, debuted in 1988 with significant contributions from Don Bluth, George Lucas, and Steven Spielberg. This franchise, consisting of an initial film followed by 13 sequels, a TV series, video games, and extensive merchandising, explores the adventures of five young dinosaurs who learn key life lessons about friendship and teamwork through their prehistoric trials. Despite the absence of the original creators in the sequels, the series has continued to captivate audiences, emphasizing themes of community and perseverance across its extensive narrative arc. Thanks to their release, we derive some subsets from their websites [48, 49].

H.4. Appreciation

Leveraging the data derived from "Curious George," "Rabbids Invasion," and "The Land Before Time," we have significantly advanced the capabilities of our story generation models. This progress has direct and impactful implications for children's education by enhancing their imaginative faculties and fostering a keen interest in learning. By integrating elements from these animated series into our models, we not only engage young minds but also deepen their affection for animated storytelling. Consequently, this not only meets but also amplifies educational objectives, such as improving literacy and cognitive skills through enjoyable and interactive content. The successful application of data from these beloved animations in our research exemplifies how academic pursuits can harmoniously blend with educational entertainment, ultimately delivering multifaceted benefits that extend well beyond conventional learning environments.

Lastly, we extend our profound appreciation to the creators and maintainers of "Curious George," "Rabbids Invasion," and "The Land Before Time," each a rich and vibrant resource that has significantly contributed to the scope and success of our research. The engaging narratives and characters from these series, especially the ever-curious George, the mischievous Rabbids, and the adventurous dinosaurs from The Land Before Time, have provided invaluable data that enhanced our narrative generation models. This project benefited immensely from the educational and entertaining content crafted with meticulous attention to detail, fostering imagination and learning in young audiences. We acknowledge the pivotal role that these animated series have played in advancing academic research aimed at educational technology. The commitment of the teams behind these beloved series to fostering curiosity and learning is both inspiring and exemplary. We are immensely grateful for the opportunity to incorporate such cherished resources into our scholarly work.

I. Broader Impacts

This project may potentially produce copyrighted content, particularly when used inappropriately or without adherence to existing intellectual property laws. To mitigate this risk, we will implement a rigorous compliance framework that respects the copyrights of third parties. This involves setting strict usage licenses that align with the legal standards dictated by our data sources. Our aim is to protect intellectual property rights while fostering innovation and ethical use of our technology. We also commit to educating users on the importance of respecting intellectual property rights when using our technology. This will be achieved through detailed user guidelines, training sessions, and readily available support to help users understand and navigate the complexities of copyright laws. By taking these measures, we aim not only to comply with legal standards but also to promote a culture of respect for intellectual property within our user community, thereby contributing positively to the broader digital ecosystem.

J. Limitations

Lack of Realistic Data Experimentation: This limitation points to a potential gap in the validation of the SEED-Story model under practical, real-world conditions. Without experiments using realistic data, it's difficult to ascertain how the model would perform in scenarios that are not perfectly controlled or that deviate from the training conditions. This can be crucial, especially in applications like storytelling where the context and variability of real-world data play significant roles. A possible solution would be to incorporate a broader range of test conditions, including noisy data or data from "in-the-wild" storytelling scenarios, to evaluate the robustness and adaptability of the model.

Training on a Non-Diverse Dataset: The second limitation is the restriction of the model's training to animation datasets which does not cover a large scale or diverse styles. This can severely limit the model's ability to generalize and produce outputs in styles that are not represented in the training data. This is particularly limiting in creative tasks such as storytelling where the ability to adapt to various artistic and narrative styles is crucial. To mitigate this, expanding the dataset to include a wider array of styles, genres, and visual aesthetics could be beneficial.

a) And indeed, it had! A small, brown monkey named George was sitting on a nearby tree branch, hugging his legs with a curious expression.
b) To his surprise, two raccoons, with mouths open and startled expressions, were looking at George. He was only partially visible, with his red hat peeking out from the bush.
c) Suddenly, a character wearing a red and white shirt, blue pants, and a yellow cap appeared. He stood a red bike with a basket, in the lush green park with trees and a wooden fence.
d) The character stood next to his red and yellow bicycle in the green park. He raised a hand to wave at George.

a) Suddenly, the man in the yellow hat stopped to look at something off-screen. George, the curious monkey, also turned to look in the same direction.
b) The bellhop continued his conversation on the phone while another figure held wooden planks. The room was filled with a classic armchair, console table with a flower vase, and a framed painting.
c) Suddenly, there was a spill! An eggs had fallen, and cracked eggs, yolk, and eggshells were scattered.
d) George sat on the floor with the man in the yellow suit and the bellhop in the red coat, amidst scattered pieces of paper and a broken object.

a) The man looked to the side with a surprised expression. The background appeared to be the interior of a rounded structure with a window.
b) A small monkey looked up at a person wearing an outfit. In the background, another person stood next to a large dinosaur skeleton inside what appeared to be a museum.
c) A dinosaur skeleton stood inside a museum, with a woman in a coat gesturing towards it. In the background, a small figure waved enthusiastically from behind a glass window.
d) An woman with short hair and a lab coat stood with hands on hips, smirking. The background showed a wall with two light switches.

Figure D. Story visualization comparison of SEED-Story and other story visualization methods.

SEED-Story

1 One day, George, the curious brown monkey, decided to explore a new room. He peeked out from behind a dresser, looking both curious and cautious. The dresser had three drawers, each with a round handle. An electrical outlet was visible on the wall.
2 George continued to explore, He looked around with a curious expression, wondering what adventures awaited him.
3 George then decided to sit on the floor, leaning against the wall next to the dresser. Suddenly, George heard a noise.
4 To his surprise, the noise was George‘s friend, a small brown dog. He take him to the hall, with the lamp lighting in the dark.
5 He stood up and looked to the right with a surprised expression. George and the dog then played a game of hide and seek. George hid behind a couch.

MM-interleaved

1 One day, George, the curious brown monkey, decided to explore a new room. He peeked out from behind a dresser, looking both curious and cautious. The dresser had three drawers, each with a round handle. An electrical outlet was visible on the wall.
2 george moved on to explore a dimly lit hallway. the walls were covered in teal paint, and there was an open door at the end of the corridor leading into another room filled with scattered objects like books and blocks.
3 suddenly, george's attention was drawn to a kitchen cabinet. his wide eyes and open mouth showed how surprised he was by the neatly arranged plates inside one of its drawers!
4 george, wearing a white bib with red spots and colorful dots on the collar, decided to make some mess. he picked up colored sand from an open booklet in front of him while sitting against yellow countertops next to various household items like bottles or containers.
5 next, george found a red ball and decided to have some fun. he jumped in the air near an open blue dresser with yellow gloves inside it on wooden floors.

Figure E. Multimodal Story generation comparison of SEED-Story and MM-Interleaved.

SEED-Story

1 Meanwhile, George stood amidst a messy bedroom with toys, books, and clothes scattered on the floor. A colorful bed with blue sheets was visible.
2 suddenly, george found himself in a colorful attic filled with paint spills. art supplies and toys were scattered on the floor as he stood amidst chaos with an expression of surprise or curiosity.
3 after cleaning up, george sat at a table with the man in yellow. they enjoyed butterfly-themed breakfast plates and drinks beside open windows showcasing cityscapes.
4 the man in yellow gestured towards a messy room with scattered toys and clothes. george looked on, holding onto his hand as they prepared for bedtime
5 as night fell, george snuggled under a blue blanket in the bed with red headboard. he looked content and happy, ready for another day of adventures tomorrow.

MM-interleaved

1 Meanwhile, George stood amidst a messy bedroom with toys, books, and clothes scattered on the floor. A colorful bed with blue sheets was visible.
2 George decided to paint. He sat on a stool, holding a paintbrush covered in pink paint. The room was messy with scattered toys, a canvas, and a drawing of an apple on the floor.
3 George then sat on the floor, holding a paintbrush with paint on it. Above him, there was a shelf with various items.
4 George sat on the floor, facing a colorful bed, in a messy room with scattered toys, a canvas, and a broken model airplane.
5 Later, George sat on the floor, smiling, with one hand on his stomach. Behind him, a bookshelf filled with colorful books stood next to a window.

Figure F. Multimodal Story generation comparison of SEED-Story and MM-Interleaved.

Figure G. Story visualization result on Rabbids Invasion data.

Figure H. Story visualization result on Rabbids Invasion data.

Figure I. Story visualization result on The Land Before Time data.

Figure J. Examples of multimodal story generation from SEED-Story. It shows two narrative branches generated from the same initial image. The top branch starts with text referencing "the man in the yellow hat," leading to images that include the character. The bottom branch starts without mentioning the man, resulting in stories that diverge from the first by excluding him.

Figure K. Multimodal long story generation results of SEED-Story.

Figure L. Multimodal story generation results of SEED-Story.