StoryCrafter: Comic-Style Storyboarding Meets 3D Camera Animation

Alex Gu
Stanford University
alexgu@stanford.edu

Isaac Kan
Stanford University
isaackan@stanford.edu

Jack Le
Stanford University
jackle@stanford.edu

Abstract

We present a novel approach to long-form text-to-video generation that combines character consistency with dynamic camera animation. Current text-to-video models struggle with maintaining character identity and visual coherence across extended narratives. Our method addresses this limitation through a three-stage pipeline: (1) generating character sheets from text descriptions or single images to establish consistent character representations, (2) tuning the diffusion model to generate character consistent anchor frames, and (3) applying a video model such as LTX-Video [4] to animate the static anchor frames into dynamic sequences. We explore pose-conditioned ControlNet [13] for generating character sheets, and StoryDiffusion [14] and low-rank adaptation with Dreambooth [10] for character-consistency. Our results demonstrate significant consistency improvements compared to baselines, preserving character identity across different scenes while enabling complex video motion for cinematic storytelling.

1. Introduction

Recent advances in generative AI have made it possible to create videos from text descriptions, as seen in models like Meta's Make-A-Video and Google's Imagen Video [11, 5]. However, these text-to-video models are limited in the length and complexity of the content they can handle: while they excel at generating short clips (a few seconds), they struggle to maintain coherence for longer narratives [3]. In fact, generating long and coherent video sequences from a lengthy text is still an open challenge, as current models cannot reliably preserve consistency of scenes or characters over multiple shots. Thus, research suggests treating a long video as a sequence of shorter "shots" (as done in cinematic storytelling) to make the task more manageable [1].

Given these limitations, we propose a modular approach to text-to-video generation that leverages the strengths of existing models while maintaining narrative and character coherence. Our method breaks a long-form textual description into smaller narrative chunks, focusing particularly on establishing character consistency through specialized character sheet generation and fine-tuning techniques before animating static anchor scenes using video models. By segmenting the narrative and generating scene-specific visuals, we ensure that each part of the story is correctly depicted, mitigating the loss of context and character identity over a long prompt.

2. Literature Review

To inform our text-to-video pipeline, we examined past works that address long-range coherence and subject consistency with modular components.

StoryDiffusion (SD) introduced a new self-attention mechanism that, together with a semantic motion predictor, improved consistency between generated images' subject identities for creating long-range videos [14]. This approach's success at decomposing narratives into temporally-linked image sequences motivates our own shot-level segmentation and the use of anchor keyframes that are animated later by a video model.

Similarly, other work by Kim et al. has shown that separating subject appearance from motion dynamics leads to high-fidelity generated videos that do not overfit to temporal patterns [8]. This decoupling principle is similar to our two-stage strategy, where DreamBooth/LoRA layers encode static character identity and downstream video models handle motion.

Finally, the VideoCrafter2 model showed that high-quality diffusion-based text-to-video generation can be achieved with limited data through targeted pre-training and lightweight adapters, which leads to our decision to build on open-source backbones (FLUX + ControlNet) and minimal fine-tuning to keep compute and data requirements low [2].

3. Problem Statement

Our task is to convert a text description into an animated video that illustrates the narrative while maintaining consistent characters. The text is in the form of a story, which may consist of multiple paragraphs containing several scenes that together form a cohesive story.

The output is a series of anchor images that exhibit strong character consistency, regardless of lighting, angle, or environment. These images can then be passed to any video model of choice to produce a short animated video composed of a sequence of one or more smaller video segments, each corresponding to a portion of the text (a narrative "chunk" or scene). The goal is for the output video to closely follow the events and scenes described in the text, bringing the story to life visually while maintaining coherence across different visual elements of the story.

3.1. Baseline

For our baseline, we use a standard text-to-image diffusion model with no modifications, and craft prompts for the purpose of generating character-consistent scene anchor frames. We include detailed character descriptions in our prompts to try to maintain consistency across different generations. Examples of such prompts can be found in Appendix B and C.

By providing detailed descriptions of the character's appearance in each prompt, this baseline attempts to guide the text-to-image diffusion model to generate similar-looking characters across different scenes. Importantly, this approach requires no additional training or model adaptation, relying solely on the text-to-image model's ability to interpret and consistently apply detailed character descriptions.

For fair comparison, we use the FLUX.1 [9] image diffusion model for this baseline as well as all of our methods.

4. Methods

Our approach to generating consistent character animations across narrative scenes involves three main components: character-specific dataset generation, model adaptation for character consistency, and camera animation for dynamic storytelling. These components work together to create a pipeline that can transform text narratives into animated videos with consistent characters.

4.1. Character-Specific Dataset Generation

In order to produce generations that are character-consistent, we must first generate a dataset of representative views of our character. These views are then used to fine-tune or prompt the diffusion models that generate our anchor frames. To do so, we create character-specific datasets on demand for each user query. We experimented with two methods of constructing character sheets.

MV-Adapter. This model [7] adapts a text-to-image Stable Diffusion XL (SDXL) model to generate multi-view consistent images from a single reference images. The adapter uses the priors of the pre-trained diffusion model to model novel 3D knowledge about the reference image.

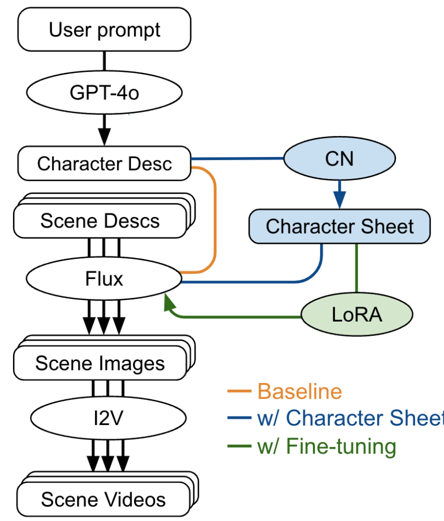

Figure 1. Overall pipeline with different diffusion methods

We found that this method worked relatively well with creating multi-view images, though there were some major drawbacks: (1) the model struggled with realistic characters, likely since the adapter was mostly trained on cartoon and anime characters, and (2) the outputs were very low-quality, which required us to add an additional upscaling step in order to get usable images for further training.

5. Character Sheet

ControlNet. This model [13] introduces additional controls to diffusion models, including depth, canny-edge, and pose. We use the pose ControlNet to condition the image generation with a pose image, which we set to be a turntable style pose image that would generate different views of the primary character. This method was very efficient and worked with a large variety of models, including more state-of-the-art image generation models such as FLUX. One disadvantage was that this method had more artifacts between different views of the character. However, we felt that the significantly faster generation time (roughly 1 minute compared to 10 minutes with MV-Adapter and upscaling) outweighed these potential inconsistencies.

With both methods, the generation process involves taking a single reference image of a character or a text description, generating a multi-view character sheet, and compiling these variations into a comprehensive character sheet that would serve as our datasets for further model tuning. This approach allows us to expand a single character reference into a diverse set of consistent representations that can be used for training character-specific diffusion models.

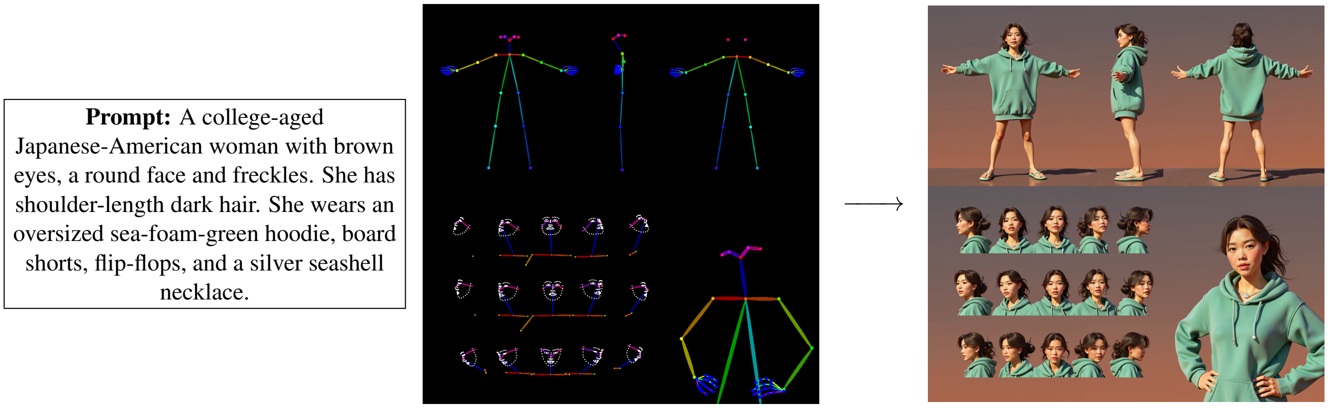

Figure 2. Example character sheet generated with ControlNet and FLUX, showing multiple views while maintaining identity consistency.

Figure 3. Character sheet generated with MV-Adapter

5.1. Model Adaptation

We explore two model architecture changes to ensure character consistency across different scenes.

Consistent Self-Attention with StoryDiffusion. To investigate inference-time character consistency, we implement Consistent Self-Attention inspired by StoryDiffusion [14]. This approach builds connections between images within a batch during the diffusion process by sampling tokens from other images and incorporating them into the self-attention mechanism. This training-free approach allows the model to maintain character identity and attire consistency across different scenes while preserving strong text controllability.

Model Fine-tuning with DreamBooth and LoRA. Using our generated character sheet as a dataset, we fine-tune a diffusion model to learn the specific visual features of the character. Our approach combines insights from DreamBooth [10] and Low-Rank Adaptation (LoRA) techniques [6]. DreamBooth teaches the model a unique identifier for the subject and associates it with the visual features of the character, while LoRA modifies a subset of the model's weights to encode character-specific information while maintaining the model's general capabilities. Both techniques allow the model to generate new images of the character in different contexts while preserving identity. This approach requires more computational resources than other inference-only techniques, though it may achieve significantly higher consistency in character representation across varied scenarios and viewpoints.

5.2. Long Range Video Generation

In order to generate long-form videos, we build on the character consistency methods to create animated sequences that maintain narrative coherence. For each narrative chunk identified in the input text, we generate representative scene images that serve as anchors (or keyframes) for our animation pipeline. These scenes include the consistent characters placed in appropriate environments as described in the text, forming a visual storyboard that guides the subsequent animation process.

Once we have generated keyframes containing consistent characters in different scenes, we employ a video model to animate the static images, concatenating the animated output of each frame to form the video. The resulting animation preserves character identity while incorporating dynamic camera work, enhancing the cinematic experience compared to the static storyboard frames. We explore the use of 2 video models to animate our anchor frames.

LTX-Video. This model [4] is a latent diffusion model that uses a Video VAE and diffusion-transformer for text-to-video generation. It is able to efficiently generate videos by leveraging a compressed latent space for the denoising objective, allowing us to animate an anchor frame into a 640×352 video within 20 seconds. In general, we are satisfied with the quality of the model's animations based on its ability to preserve character design from the initial frame.

However, we noticed that the output of the LTX-Video model is highly dependent on the animation prompt (i.e., the amount and direction of motion in the generated video depends on the camera motions described in the prompt).

Wan 2.1. This suite of open-source diffusion-transformer video foundation models (offered in 1.3B and 14B parameter variants) couples a high-capacity UNet with a 1080p video-VAE [12]. It supports multiple conditioning modes, including text-to-video (T2V), image-to-video (I2V), and first-last-frame-to-video (FLF2V). We adopt the I2V path of the 1.3B parameter model to animate anchor frames into 4s clips at 720p resolution. Compared with LTX-Video, we find that Wan 2.1 delivered higher appearance fidelity and smoother, more realistic motion. However, this came at the expense of higher compute costs and time ($\approx 2\times$ LTX-Video on our hardware). In practice, we thus default to LTX-Video for all anchor-frame animations, using Wan 2.1 only as a proof of concept.

6. Results and Analysis

6.1. Character Sheet Generation

We tried implemented the character sheet generation pipeline using both ControlNet and MV-adapater. The generated character sheets show good results in maintaining character identity across different views, angles, expressions, and lighting. Figure 2 shows an example of a generated character sheet using ControlNet, demonstrating the consistency of the character's identity despite variations in pose and perspective. The girl's face and outfit are consistent across the various perspectives, and the only difference is minor detailing in the clothes (the jean shorts are messed up on the bottom left view, and the designs on the green jacket vary slightly across the views). MV-Adapater, on the other hand, results in a more expressive range of the character in terms of facial expressions, lighting, and more. An example can be seen in Figure 3. However, we found MV-Adapter to be much more restrictive in flexibility to modify the character appearance. Practically, running MV-adapter also was significantly slower (1̃0 min) and would occasionally hang and crash.

Because of ControlNet's more flexible character customization, faster generation time, and ability to generate realistic characters, we decided to proceed with ControlNet over MV-adapter for our character sheet generation step in the pipeline.

6.2. Comic Qualitative Evaluation

Style Consistency. The baseline model produces images with noticeable variation in visual style, ranging from painterly to cartoony, with inconsistent lighting, color palettes, and rendering techniques (Figure 4). This leads to a lack of cohesion across scenes, making it difficult to read the images as part of the same narrative world. In contrast, the LoRA-Dreambooth fine-tuned model generates images with a unified hyperrealistic aesthetic. The scenes generally share a consistent tone, contrast, palette, and rendering style, creating the sense of a coherent visual universe and improving narrative flow. The StoryDiffusion comic also has a consistent realistic style throughout.

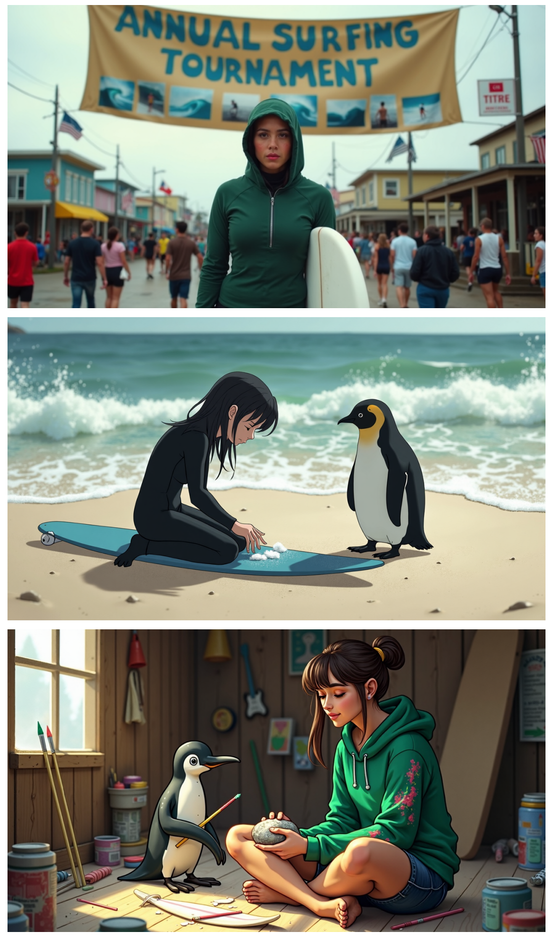

Figure 4. Frames from baseline showing distinct styles and characters

Character Consistency. The baseline model struggles to maintain a stable character identity across images. The girl's hair length, facial features, and outfit change noticeably, and the penguin companion varies in design. Notably, we see a different girl's face in every scene, likely since it wasn't specified in the character description (Figure 4). The LoRA model, however, establishes a recognizable character with consistent physical traits and a recurring, visually stable penguin companion. The girl has a consistent facial structure, hair length, and clothing style. It is extremely difficult to find differences between the girl between the scenes. The StoryDiffusion comic also has a relatively consistent character, with similar face and hair. However, the character's outfit varies significantly, with purple stripes emerging in some of the scenes, and the jacket changes color. This is likely because StoryDiffusion's reference image focuses on the character's face, rather than the outfit. Thus, the outfit is only preserved through the text character description or attention between scenes. Another practical limitation of StoryDiffusion is that prompts (including the character description) are limited to 77 tokens, so the character description must be extremely concise.

Overall Comic Quality. The images generated by the baseline approach are high quality and illustrate the story arc well. The scenes are expressive and diverse. The LoRA comic is similarly high quality. However, the StoryDiffusion-generated comic is noticeably worse. The scenes focus on the main character and leave out important parts of the plot, such as the penguin or the surfboard. The image also has significant distortions, especially in the main character's face.

Figure 5. Selected frames from StoryDiffusion comic that lack a surfboard, the key element of the plot

Resolution vs Video Quality. At inference time for video generation, we observed that the resolution at which the anchor frames were animated had an outsized impact on the quality of the generated animations. Specifically, generating videos at higher resolutions (i.e., matching the 720p resolution of the input frames) resulted in higher-quality videos. In particular, fine-grained motion suffers during low-resolution (360p) generation, with the final video degrading hand motions and adding random movement of facial structures (mouth/nose) and hair. When we increase the resolution back to 720p, these artifacts disappear, but at the expense of compute time (20s vs 90s). Our hypothesis for this behavior is that lower-resolution generation requires additional downsampling and interpolation between resolutions, affecting the quality of the animated video.

6.3. Generation time

The generation times of the various approaches can be found in Table 1. StoryDiffusion has the fastest inference image generation time, likely since it uses the older SDXL model rather than FLUX. However, it requires a initial character sheet generation overhead. Similarly, LoRA requires the character sheet as well as model fine-tuning overhead. This significant overhead can be justified by longer comic generation tasks, or by reusing characters for different comic generations.

Feature	**Baseline**	**LoRA**	**SD**
Character consistency		✓	✓
Style consistency		✓	✓
Outfit consistency	✓	✓	
Prompt consistency	✓	✓	
Overall comic quality	✓	✓	
Pretrain time (/character)	N/A	15m	5m
Inference time (/image)	20s	20s	4s

Table 1. Feature comparison across methods

7. Conclusion

We introduced a modular pipeline for long-form text-to-video generation that addresses a key limitation in current models: maintaining consistent character identity and visual coherence across extended narratives. Our method integrates character sheet generation, model adaptation for character consistency, and dynamic video animation. Among the approaches tested, LoRA fine-tuning consistently produced the highest quality results in both character and style consistency. While LoRA incurs greater overhead due to model tuning and character sheet generation, this cost is justified when the same character is reused across multiple scenes or projects. Compared to baseline and inference-only methods like StoryDiffusion, our LoRA-based approach enables more cinematic, coherent, and visu-

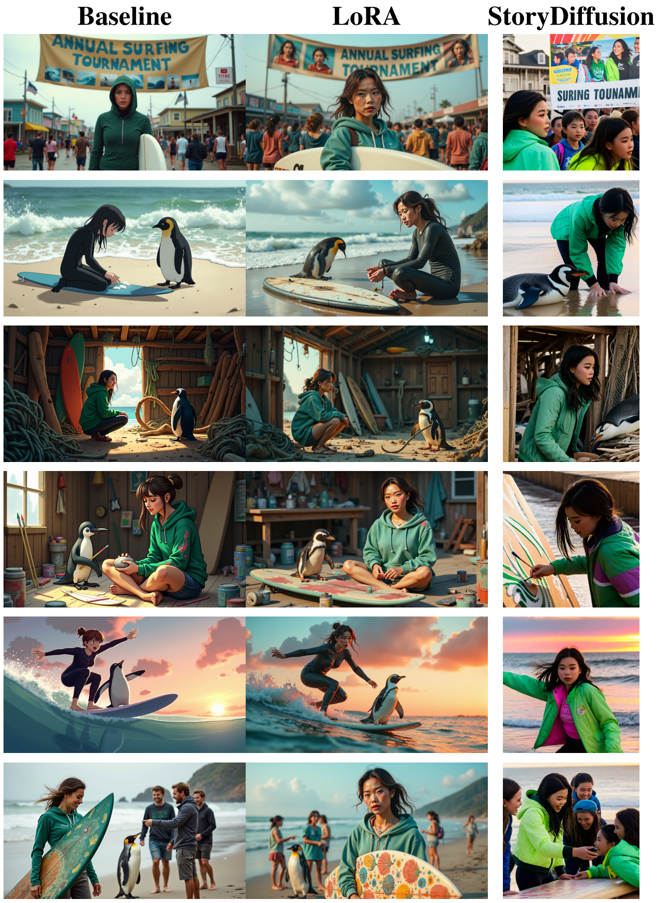

Figure 6. Comparison of selected scenes from full surfing story in Appendix 7

ally stable storytelling, pushing forward the capabilities of generative video models for narrative applications. Lastly, the parallelizability of our animation pipeline is a unique advantage granted by the anchor frame $\rightarrow$ animation approach, allowing for the efficient generation of long-form videos that is unmatched by current video models.

7.1. Future Work

There are several directions for improving and extending our pipeline:

1. **Multi-Character Consistency:** Our current system focuses on maintaining consistency for a single primary character. Future work could generalize the approach to support multiple interacting characters with consistent identities, poses, and relationships across scenes.

2. **Dialogue and Lip Sync Integration:** Adding speech generation and synchronizing character lip movement with dialogue would enhance realism. This would require aligning text-to-audio and image-to-video systems, possibly via phoneme-aware animation models.

3. **Interactive or Conditional Generation:** Enabling user feedback loops or real-time editing of characters and scenes could allow for more controllable and iterative video generation workflows, useful for creators and storytellers.

These directions aim to push beyond short-form or single-scene generation and toward robust, high-quality, and fully automated long-form narrative video synthesis.

References

[1] Y. Atzmon, R. Gal, Y. Tewel, Y. Kasten, and G. Chechik. Motion by queries: Identity-motion trade-offs in text-to-video generation, 2025.

[2] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

[3] A. Dirik. Text-to-video: The task, challenges and the current state, May 2023.

[4] Y. HaCohen, N. Chiprut, B. Brazowski, D. Shalem, D. Moshe, E. Richardson, E. Levin, G. Shiran, N. Zabari, O. Gordon, P. Panet, S. Weissbuch, V. Kulikov, Y. Bitterman, Z. Melumian, and O. Bibi. Ltx-video: Realtime video latent diffusion, 2024.

[5] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, and T. Salimans. Imagen video: High definition video generation with diffusion models, 2022.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021.

[7] Z. Huang, Y.-C. Guo, H. Wang, R. Yi, L. Ma, Y.-P. Cao, and L. Sheng. Mv-adapter: Multi-view consistent image generation made easy, 2024.

[8] D. Kim, J. Zhang, W. Jin, S. Cho, Q. Dai, J. Park, and C. Luo. Subject-driven video generation via disentangled identity and motion, 2025.

[9] B. F. Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[10] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023.

[11] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman. Make-a-video: Text-to-video generation without text-video data, 2022.

[12] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, J. Zeng, J. Wang, J. Zhang, J. Zhou, J. Wang, J. Chen, K. Zhu, K. Zhao, K. Yan, L. Huang, M. Feng, N. Zhang, P. Li, P. Wu, R. Chu, R. Feng, S. Zhang, S. Sun, T. Fang, T. Wang, T. Gui, T. Weng, T. Shen, W. Lin, W. Wang, W. Wang, W. Zhou, W. Wang, W. Shen, W. Yu, X. Shi, X. Huang, X. Xu, Y. Kou, Y. Lv, Y. Li, Y. Liu,

Y. Wang, Y. Zhang, Y. Huang, Y. Li, Y. Wu, Y. Liu, Y. Pan, Y. Zheng, Y. Hong, Y. Shi, Y. Feng, Z. Jiang, Z. Han, Z.-F. Wu, and Z. Liu. Wan: Open and advanced large-scale video generative models, 2025.

[13] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[14] Y. Zhou, D. Zhou, M.-M. Cheng, J. Feng, and Q. Hou. Storydiffusion: Consistent self-attention for long-range image and video generation, 2024.

A. Original Movie Prompt

I want to generate a series of 10 scene descriptions for a movie. I will put the description of what I want below. Please come up with 10 key scenes in the plot, then describe each scene in detail (such that I could render an image using your description), using character [x] to refer to the main character. Then, come up with a detailed character description afterwards, specifying just the character's appearance (clothes, style, etc), so that one could render what the character looks like. Here is the user description: "a pixar style coming of age story of a girl who gets a chance to compete at a surf tournament but gets smashed by a massive wave, it is only after she makes her own board with a fat penguin that she realizes the true meaning of surfing is to have fun"

B. Generated Scene Descriptions

1. **Scene: Surfing Dreams** - The sun rises over a sparkling ocean, casting a warm orange glow over the beach. [x], a spirited girl with tousled hair, stands atop a sand dune, clutching her old, beat-up surfboard. In the distance, elite surfers carve through waves, their sleek boards slicing through the water. [x]'s eyes shimmer with ambition.

2. **Scene: Announcement of the Big Tournament** - The town square is abuzz with excitement as a giant banner unfurls, announcing the annual surfing tournament. [x] and her friends crane their necks to see the poster featuring glossy images of past champions. In the corner, a group of pro surfers smirk, sizing up [x] and her battered board.

3. **Scene: The Big Wave** - The ocean churns violently under a stormy sky. [x] paddles out nervously, eyes wide as she approaches a monstrous wave. The wave rises higher, casting a shadow over her. [x] takes a deep breath and drops in, but the board wobbles and slips out from under her. The wave crashes down, swallowing her in a furious, frothing barrel.

4. **Scene: Washed Ashore** - The aftermath of the wipeout. [x] lies sprawled on the sand, coughing up water as her broken board washes up beside her in splinters. Seagulls circle overhead. In the background, a rotund, goofy-looking penguin with a perpetually blank expression waddles up, inspecting the wreckage.

5. **Scene: The Penguin's Lair** - Inside a cozy shack filled with surfboards and fishing nets, the penguin drags in scraps of driftwood and random beach trash. [x], still drenched and dazed, watches as the penguin comically attempts to assemble a surfboard out of the wreckage. Despite its clumsiness, the penguin shows unexpected craftsmanship.

6. **Scene: Building the New Board** - Sunlight filters through the shack's wooden slats, illuminating [x] and the penguin as they work together to shape and sand a new surfboard. The penguin uses its flippers to smooth the edges, and [x] paints colorful patterns on the board, inspired by ocean waves and penguin feathers. They share a laugh, covered in sawdust and paint.

7. **Scene: First Ride on the New Board** - The water is calm and shimmering under a pastel sunset. [x], now on her quirky, hand-made board, paddles out with the penguin balanced precariously on the nose. They catch a small wave, riding it together as the penguin flaps its wings excitedly, wobbling but grinning.

8. **Scene: Reconnecting with Friends** - Back on the beach, [x] reunites with her old surfing friends, who stare in awe at her new board and the chubby penguin sidekick. The atmosphere is warm and friendly as they all share waves, taking turns riding the board, laughing and wiping out together.

9. **Scene: The Tournament Revisited** - The tournament is in full swing, with pro surfers executing flawless tricks on sleek, shiny boards. [x] steps up to the starting line, her hand-made board under her arm and the penguin waddling beside her. The crowd whispers, some amused, some skeptical. [x] takes a deep breath, eyes focused on the horizon.

10. **Scene: Finding the Fun** - A massive wave rises, and [x] paddles in, catching it with the penguin clinging to her back. Instead of trying to dominate the wave, she dances playfully, allowing the board to carve fun, unpredictable patterns. The crowd watches, first in confusion, then in delight, as [x] laughs and the penguin flaps its wings, celebrating the sheer joy of the ride.

C. Character Description

A college-aged Japanese-American woman with large expressive brown eyes and long dark lashes, round face with a slight angular jaw, sun-kissed skin peppered with freckles across cheeks and nose, thick, well-shaped eyebrows showing quiet determination, wavy, shoulder-length dark-brown hair, tousled and tangled from the sea with a few sun-bleached strands. She wears an oversized, worn sea-foam-green hoodie with frayed cuffs over a snug, patched wetsuit that sports colorful, hand-stitched repairs; loose board shorts; silver seashell necklace resting at her collarbone, in flip-flops.

D. Sample comic comparisons

Figure 7. Comparison between methods of generating a surfing story

Figure 8. Comparison between Baseline and LoRA for John Wick story

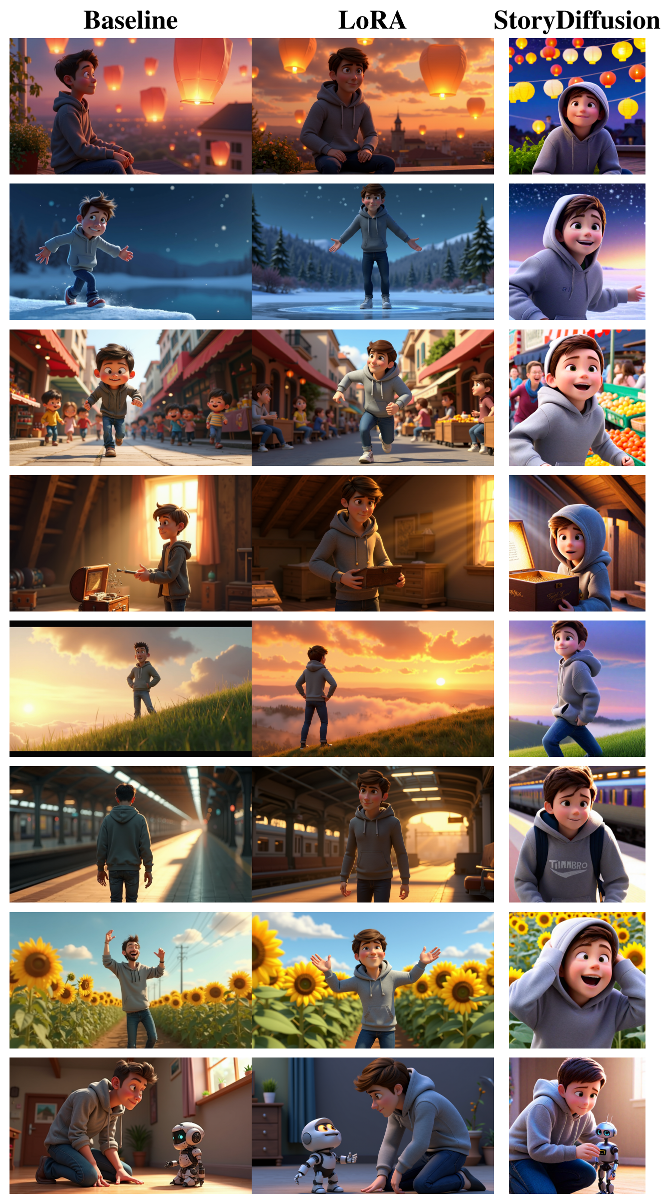

Figure 9. Comparison between methods of generating a Timbro story