

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

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 by breaking a long-form textual description into smaller narrative chunks. Our approach focuses particularly on establishing character consistency through specialized character sheet generation and fine-tuning techniques before animating static anchor scenes using video models such as Stable Virtual Camera [13] or LTX-Video [4]. 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 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 long-form text description into a short animated video that illustrates the narrative while maintaining consistent character representations. The text narrative will be in the form of a story or descriptive passage, which may consist of multiple paragraphs containing several distinct scenes or events that together form a cohesive story. Alternatively, the input could include a single image of a character along with a text description of scenes where the character should appear.

The output is a short animated video that visualizes the given narrative with consistent character representations, regardless of lighting, angle, or environment. The video is 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 are included in Appendix C and D.

By providing detailed textual 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 following 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.

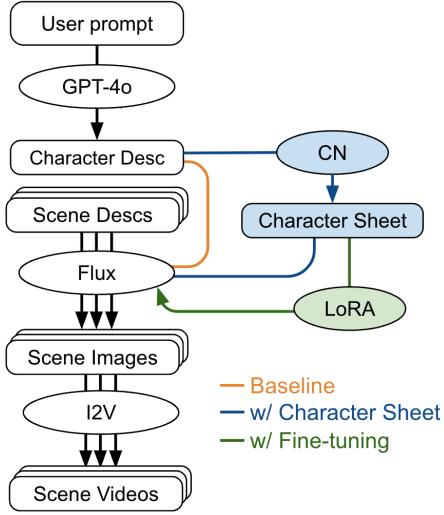


Figure 1. Overall pipeline with different diffusion methods

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. 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 [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. 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

ControlNet [12] 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 turn-table 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.

4.2. Model Adaptation

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

4.2.1 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.

4.2.2 Model Fine-tuning with DreamBooth and LoRA

We further explore model fine-tuning techniques. 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

Prompt: A cat standing upright wearing a chef's uniform, hyper-realistic oil painting, character sheet, in different poses and angles, including front view, side view, and back view, turnaround sheet, floating in outer space background, visible face, portrait

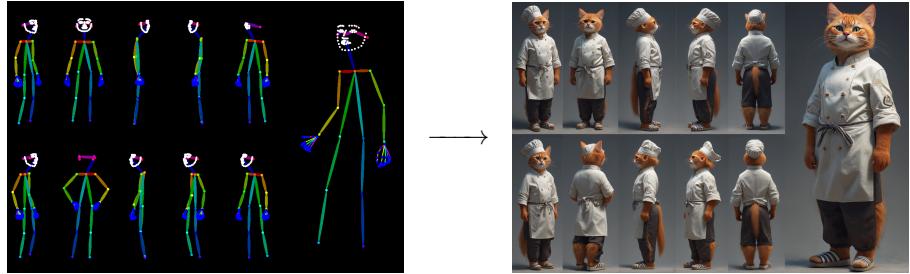


Figure 2. Example character sheet generated with ControlNet and FLUX, showing multiple views while maintaining identity consistency.

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.

4.3. 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 video models such as Stable Virtual Camera [13] and LTX-Video [4] to animate the static images. The resulting animations preserve character identity while incorporating dynamic camera work that enhances the narrative flow, creating a cinematic experience from the static storyboard frames.

5. Preliminary Results

5.1. Character Sheet Generation

We have successfully implemented the character sheet generation pipeline using both ControlNet and MV-Adapter. The generated character sheets show promising 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. Appendix E contains the character sheet for MV-Adapter as well as further discussion on the method.

5.2. Quantitative Evaluation

We evaluate our character-consistent scene generation methods using image-image CLIP scores between scenes. Table 1 shows preliminary results with the baseline images from Appendix A. The CLIP score was used in [14] to measure character similarity across scenes, and we use it for the same purpose here.

Method	CLIP Score
Prompt Engineering	0.8401
Story Diffusion [14]	0.8950
Character Sheet + Fine-tuning	x

Table 1. Quantitative evaluation of character similarity

Our preliminary results show that the prompt engineering baseline falls short of the reported character similarity scores from [14], indicating that StoryDiffusion demonstrates superior character consistency in generation.

5.3. Future Work

We aim to complete the following tasks for our project:

1. Completing the implementation of the remaining character consistency approaches, analyzing the trade-offs between model complexity and consistency quality.
2. Fully integrating our character generation pipeline with Stable Virtual Camera or other video models to create an end-to-end system for animated storyboards.
3. Integrating additional quantitative perceptual metrics, such as LPIPS and CLIPSIM, and running small human evaluation studies to get better qualitative evaluations.
4. Optimizing the transition mechanisms between animated segments to create smoother story flow. Our method currently just has sharp cuts between narrative chunks.

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] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models, 2023.
- [13] J. Zhou, H. Gao, V. Voleti, A. Vasishta, C.-H. Yao, M. Boss, P. Torr, C. Rupprecht, and V. Jampani. Stable virtual camera: Generative view synthesis with diffusion models, 2025.
- [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. Baseline Scene Anchors



Figure 3. Baseline scene images generated using FLUX.1 with the prompts from Appendix C

B. 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"

C. Generated Scene Descriptions

- 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.
- 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.
- 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.
- 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.
- 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.
- 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.

- 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.
- 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.
- 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.
- 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.

D. Character Description

Mia Watanabe is a college-aged Japanese-American girl with a Pixar-style design, featuring large, expressive brown eyes framed by long, dark lashes. Her face is round with a slight angular jawline, and her skin is sun-kissed, dotted with freckles across her cheeks and nose.

She has wavy, dark brown hair that falls just past her shoulders, often tangled and tousled from the sea, with a few lighter strands hinting at sun exposure. Her eyebrows are thick but well-shaped, conveying a sense of determination and intensity.

Mia typically wears a worn, oversized hoodie in a seafoam green hue, its sleeves slightly frayed. Underneath, she wears a snug, patched-up wetsuit with colorful, hand-stitched sections that hint at her artistic, resourceful nature. Around her neck hangs a silver seashell necklace that she never takes off — her only connection to her late mother.

She walks with a confident stride, her board shorts slightly oversized, hanging loosely around her hips. Her feet are usually bare or in well-worn flip-flops, toes slightly sandy and tanned.

Prompt:

A 3D Pixar-style animated character of a college student with exaggerated proportions and cartoonish features. She should have a slightly oversized head, large expressive eyes, and a button nose. The character stands with arms outstretched, wearing a cozy, slightly baggy hoodie with some texture, blue jeans, and colorful sneakers. Her facial expression is neutral with the signature Pixar warmth and subtle emotion. The lighting is bright with soft shadows, and the rendering style should match Pixar's aesthetic.



Figure 4. Example character sheet generated using MV-Adapter.

E. MV-Adapter Character Sheet

Figure 4 shows an example character sheet generated using the MV-Adapter method, using the SDXL text-to-image model, which is the most high quality model that MV-Adapter supports. As seen, the sheet shows multiple views and character expressions of the character while maintaining identity consistency.

In our testing so far, we've found that MV-Adapter performs best on human characters with an animated style. It was unable to model non-human characters, such as the chef cat shown in Figure 2. Furthermore, a single run of the MV-Adapter takes over 10 minutes, due to the various upscaling steps necessary due to the lower expressivity of the SDXL model.