Long Context Tuning for Video Generation

Supplementary Material

A. Implementation Details

A.1. Training

We train the 3B parameter MMDiT [3] video diffusion model with 128 NVIDIA-H800-80GB GPUs. We maintain a balanced $1:1$ mixing ratio between pre-training single-shot data and our curated scene-level video dataset. We preserve the original aspect ratio of videos without cropping [2], only downsampling to achieve a resolution size equivalent to 480×480 pixels. All training videos (including individual shots within each scene) are sampled at 12 frames per second and limited to a maximum duration of 125 frames (approximately 10 seconds). For scene-level video data, we implement a random frame substitution where every shot has a 0.1 probability of being replaced by a randomly selected single frame from that same shot, thereby enabling image conditioning and generation capabilities. To enhance model flexibility and generalization, we randomly omit the story overview from the global prompt during training. Our shot sampling strategy from scene video batches employs a biased probability distribution that favors sampling as many shots as possible while still accommodating smaller shot counts, effectively supporting dynamic context window sizes. We exclude the dummy video tokens of the global prompt from loss computation to focus training on meaningful content.

For distributed training, we implement PyTorch's Fully Sharded Data Parallel (FSDP) framework. We leverage sequence parallelism to distribute token sequences across four devices, complemented by gradient checkpointing techniques to reduce GPU memory usage. We employ the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay coefficient of 0.1. The learning rate is initialized at 1×10^{-4} and modulated using a cosine decay schedule. These optimizations collectively enable a substantial per-GPU batch size of 3.5 video sequences. We apply Exponential Moving Average (EMA) to the weights with a decay factor of 0.996.

A.2. Inference

In our experimental evaluation, we discover that multi-shot generation requires no specialized inference configurations beyond those used for single-shot generation, allowing for a unified inference approach. Specifically, we employ a 32-step Euler sampling algorithm to progressively transform Gaussian noise into coherent visual content. We implement classifier-free guidance exclusively on textual conditions with a guidance scale of 7.5 to balance creativity and fidelity. To address the inherent differences in detail levels between global narrative prompts and shot-specific descriptions, we employ differentiated negative prompting strategies: comprehensive negative prompts for global contexts and concise negative prompts for per-shot instructions. This specialized approach significantly enhances visual quality and consistency across generated shots.

B. Baseline Adaptation

VideoStudio [6]. VideoStudio incorporates additional reference appearance embeddings to maintain identity consistency across multiple shots. To implement this approach, we utilize FLUX [1] to synthesize character and environment images based on descriptions generated by o3-mini [7]. Subsequently, we provide these appearance images as conditions for shot generation, strategically selecting references based on their semantic relationships to the current shot.

StoryDiffusion [9]. StoryDiffusion ensures visual consistency by leveraging attention features from previously generated shots. In our implementation, we first generate core semantic elements, such as characters and environments, and then reference these elements when producing subsequent shots. To enable multi-identity generation capabilities, we modified the general prompt structure in the official implementation to specifically describe the major character of each shot, rather than applying the same generic prompt across all shots.

In-Context LoRA [4]. In-Context LoRA concatenates all shot descriptions into a single unified prompt and generates a 3×1 image grid. To adapt our content to this specific prompt format, we provide exemplars from the official implementation to o3-mini as a reference, allowing the language model to transform our descriptions into the required format. During our experiments, we observed that In-Context LoRA occasionally produces imperfect image grids with excessively large borders, necessitating multiple generation attempts to achieve optimal results.

Kling [5]. We employ Kling, a state-of-the-art commercial video generation model, for image-to-video animation of keyframe-based methods such as VideoStudio and StoryDiffusion. We utilize Kling's default generation settings, supplying the keyframes as initial frames and conditioning the generation with translated Chinese prompts to optimize performance.

C. Prompts

LLM Director Prompts. We observe that large language models (LLMs) excel at narrative construction and video script development. Consequently, we employ o3-mini [7] to craft sophisticated video scripts for our evaluations and demonstrations. In Tab. 1, we present the comprehensive prompt used to instruct the LLM to function as a movie director, guiding it to design cohesive scenes with well-defined characters, distinctive environments, and compelling story descriptions.

Multi-shot Generation Prompts. We instruct o3-mini [7] with the prompts in Tab. 1 to generate multiple lengthy and detailed video scripts for our multi-shot generation evaluations. For brevity, we omit the complete prompts in the main paper. In Tab. 2, we present the complete prompt used for the qualitative comparison featured in the main paper.

Compositional Generation Prompts. For compositional generation, we condition the model to generate new shots using both an identity image and an environment image as inputs. We leverage Gemini [8] to generate detailed descriptions of the condition images, which we incorporate into the *global* prompt, while maintaining concise *per-shot* prompts. We apply noise at $t = 100$ to the condition images. In Tab. 3, we provide the complete prompts for the compositional generation example in the main paper.

D. Computational Overhead

In Fig. 1, we analyze LCT's inference latency and computational requirements across different shot counts (x-axis), benchmarked against a conventional independent single-shot T2V baseline (which cannot maintain inter-shot consistency). Although our method delivers superior scene consistency with some additional computational cost, our implementation of a causal architecture with KV-cache substantially mitigates this overhead, making the approach practical for multi-shot video generation.

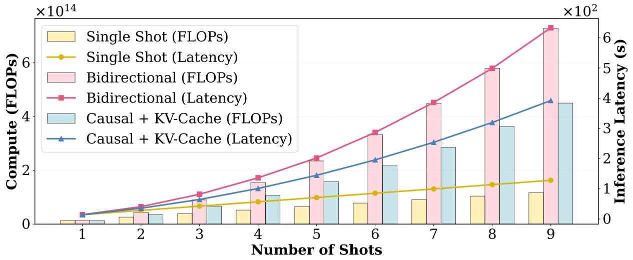

Figure 1. **Computational Overhead Analysis.** The causal model with KV-cache significantly reduces the inference computation.

Table 1. Instructions LLM Director.

You are now a movie director writing scripts for movie scenes with multiple shots. A script must contain a global description and several per-shot descriptions. The global description details shared elements across shots, including characters, environment, and storyline overview. Per-shot descriptions detail what happens in each shot. Characters in her per-shot descriptions must be referred to using “Character [ID num.]”, *e.g.*, “Character 1”, “Character 2”, avoiding pronouns or ambiguous references like “he”, “she”, “they”, “two characters”, “both characters”. Here’s an example:

A. Global Description
A.1 Character
Character 1: A young man with short black hair, wearing a gray sweater vest, a white collared shirt, a striped tie, and a blue apron.
Character 2: An older man with short graying hair, wearing a dark suit jacket, a white collared shirt, a patterned tie, and a blue apron.
A.2 Environment
The scene takes place within a brightly lit bookstore. Wooden bookshelves line the walls, stocked with various books. Festive decorations, including a Christmas tree adorned with lights, suggest the holiday season. A ”Staff Only” sign is visible on a frosted glass door. The overall ambiance is warm and inviting.
A.3 Storyline
Inside a bookstore, Character 2 shares a personal anecdote about his divorce with Character 1. He expresses the coldness he feels post-divorce and a wistful hope for his wife’s return. He mentions a philosophical point about destiny. Character 1 listens attentively and reacts with a smile and subtle head movements, showing empathy. Then, the focus shifts to a young girl browsing books.

B. Per-shot Description
Shot 1. Character 2 leans on a table and talks to Character 1. Character 1 stands opposite, listening and occasionally glancing away. Character 2 initially looks down, then turns to Character 1, making eye contact. The shot frames them from the waist up within the bookstore setting.
Shot 2. This shot focuses solely on Character 1, who stands and continues to speak, still facing where Character 2 is standing.
Shot 3. The camera cuts to a different part of the bookstore, showing a young girl with two braids browsing through manga books displayed on a table. She is wearing a beige sweater, blue overalls, and beige shoes. There is a store promotion poster and the bookstore aisle in the foreground.
Shot 4. . . .

Based on the requirements and example format, please design a story and write the script for it. You need to follow the format in the example. The script should contain 9 shots in total.

Table 2. Complete Prompts for Qualitative Comparison.

Global Prompt	Character 1: A woman in her early 30s with shoulder-length auburn hair, dressed in a chic dark coat layered over a patterned blouse, paired with tailored jeans and ankle boots. Character 2: A man in his early 40s with salt-and-pepper hair, wearing a light brown blazer over a crisp white shirt, dark trousers, and polished loafers. The scene is set in a cozy urban cafe featuring exposed brick walls and wooden tables. Warm ambient lighting emanates from hanging pendant lights, casting inviting shadows over vintage posters and a chalkboard menu. Small potted plants and the gentle hum of conversation combine with soft jazz playing in the background to create an intimate and relaxed atmosphere. Character 1 is enjoying a quiet moment by a window when Character 2 enters the cafe. Their paths cross as Character 2 orders a coffee and casually notices Character 1. A spontaneous conversation ensues, during which both characters reveal fragments of past regrets and dreams of new beginnings. Gradually, their initial glances evolve into a meaningful exchange, suggesting the start of a deep, personal connection.
Shot 1	A wide establishing shot of the cafe's interior view. The camera pans slowly across wooden tables, vintage decor, and warm lighting, setting the stage for the encounter.
Shot 2	A medium shot of Character 1 seated by a large window. Character 1 is absorbed in a book while gently sipping coffee, with natural light accentuating a quiet, reflective expression.
Shot 3	A close-up of the coffee and book in Character 1's hand.
Shot 4	A shot captures Character 2 entering the cafe. The doorbell chimes, and Character 2 pauses for a moment, taking in the inviting ambiance.
Shot 5	A close-up on Character 1 reveals a subtle shift in expression as Character 1 glances sideways toward the entrance, showing the first spark of curiosity.
Shot 6	An over-the-shoulder shot from behind Character 2 shows Character 2 approaching the counter to order coffee. The camera briefly captures a glimpse of Character 1 in the background, seated near the window.
Shot 7	A shot capturing Character 2 walking towards Character 1 and sitting beside, breaking the silence with an inviting conversation opener.
Shot 8	A close-up captures Character 1 listening intently. The camera emphasizes Character 1's empathetic eyes and subtle nods, reinforcing the deepening of the conversation.
Shot 9	A medium two-shot of Character 1 and Character 2 seated at adjacent tables. The scene shows their quiet conversation as Character 2 makes a friendly gesture toward Character 1.
Shot 10	A close-up focuses on Character 2's expressive face. Character 2 speaks softly, with eyes reflecting hints of nostalgia and a touch of vulnerability as personal memories surface.
Shot 11	A wide shot showing both Character 1 and Character 2 as they engage in the conversation, revealing the cafe interior surrounding them.

Table 3. Complete Prompts for Compositional Generation.

Global Prompt	Character 1: an older man with short, slightly disheveled gray hair. He has fair skin, a slightly lined face, and blue eyes. He is wearing a white, checkered button-down shirt under a black V-neck sweater. The scene is set in the interior of an art museum. The space is characterized by its polished herringbone-patterned wooden floor and white or light-colored walls. Along the walls, a variety of framed paintings are displayed, ranging in size, subject matter, and style. The frames themselves are ornate and varied in color, adding to the visual interest. The lighting appears to be a combination of natural and artificial, with a series of track lighting fixtures installed on the ceiling. The ceiling is high and features recessed panels and other architectural details. The overall atmosphere is one of quiet sophistication and cultural richness, inviting contemplation and appreciation of the artworks on display. Character 1 is drinking coffee in the museum.
Identity	A close-up of Character 1 with greenery and a building in the background.
Env.	A wide shot of the art museum, showing its interior layouts.
Output	Character 1 is drinking coffee in the museum.

References

[1] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1

[2] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023. 1

[3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1

[4] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024. 1

[5] Kuaishou. Kling video model. https://kling.kuaishou.com/en, 2024. 1

[6] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videostudio: Generating consistent-content and multi-scene videos. In *European Conference on Computer Vision*, pages 468–485. Springer, 2024. 1

[7] OpenAI. o3-mini. https://openai.com/index/openai-o3-mini/, 2024. 1, 2

[8] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[9] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024. 1