

# Long Context Tuning for Video Generation

Yuwei Guo<sup>1,2</sup> Ceyuan Yang<sup>2,†</sup> Ziyan Yang<sup>2</sup> Zhibei Ma<sup>2</sup> Zhijie Lin<sup>2</sup>  
 Zhenheng Yang<sup>3</sup> Dahua Lin<sup>1</sup> Lu Jiang<sup>2</sup>

<sup>1</sup>The Chinese University of Hong Kong <sup>2</sup>ByteDance Seed <sup>3</sup>ByteDance

## Abstract

Recent advances in video generation can produce realistic, minute-long single-shot videos with scalable diffusion transformers. However, real-world narrative videos require multi-shot scenes with visual and dynamic consistency across shots. In this work, we introduce Long Context Tuning (LCT), a training paradigm that expands the context window of pre-trained single-shot video diffusion models to learn scene-level consistency directly from data. Our method expands full attention mechanisms from individual shots to encompass all shots within a scene, incorporating interleaved 3D position embedding and an asynchronous noise strategy, enabling both joint and auto-regressive shot generation without additional parameters. Models with bidirectional attention after LCT can further be fine-tuned with context-causal attention, facilitating auto-regressive generation with efficient KV-cache. Experiments demonstrate single-shot models after LCT can produce coherent multi-shot scenes and exhibit emerging capabilities, including compositional generation and interactive shot extension, paving the way for more practical visual content creation. See our [Project Page](#) for more details.

## 1. Introduction

Video generation has experienced significant advancements in recent years. By leveraging web-scale data and scalable model architecture such as diffusion transformer (DiT) [40], state-of-the-art models (*e.g.*, SoRA [7], Kling [32], Gen3 [44]) are now capable of synthesizing realistic single-shot videos lasting up to a minute. However, real-world narrative videos, such as movies and television shows, are composed of multiple single-shot segments. This implies a substantial gap between the current capabilities of single-shot video generation and the authentic demands of video content production.

To bridge this gap, video generation research may need

to evolve from single-shot synthesis to scene-level generation. In this higher-level paradigm, a scene is defined as a series of single-shot videos capturing coherent events unfolding over time [9, 42, 43]. For instance, the classic scene in *Titanic* where *Jack* and *Rose* meet on the deck consists of four principal shots (Fig. 2): (1) close-shot of *Jack* looking back, (2) mid-shot of *Rose* speaking, (3) wide-shot of *Rose* approaching *Jack*, and (4) close-up shot of *Jack* embracing *Rose* from behind. Generating such a scene requires consistency in *visual appearance* and *temporal dynamics* to ensure a coherent narration flow. Specifically, visual coherence indicates the consistent rendering of shared visual components, *e.g.*, person identity, background, lighting, and color tone. Similarly, temporal coherence necessitates the uniformity in dynamic elements such as character actions (*e.g.*, maintaining consistent walking pace) and camera movements (*e.g.*, smooth versus shaky camera) across shots. Addressing these challenges demands novel approaches for scene-level video generation.

Many solutions have been proposed to address scene-level video generation, most of which fall into two categories: (1) appearance-conditioned generation [27, 29, 35], and (2) keyframe generation [26, 73, 74, 76] followed by image-to-video (I2V) animation [5, 61]. In the first paradigm, key visual elements (*e.g.*, character identity and background) serve as conditional inputs to enforce cross-shot consistency. However, this approach struggles to maintain abstract elements like lighting and color tone due to its reliance on predefined conditions and specially curated datasets. The keyframe-based strategy generates a coherent set of keyframes to ensure visual consistency, which then serve as initial frames for an I2V model to synthesize each shot independently. Yet, the independent shot synthesis fails to guarantee temporal consistency across shots. Moreover, sparse keyframes limit the effectiveness of conditioning. For instance, if a character enters the scene between keyframes, the keyframe-based method misses this character as part of the conditioning, thereby rendering a consistent representation of the character is infeasible.

In this work, we propose *Long Context Tuning (LCT)* for scene-level video generation. LCT builds upon a pre-

---

<sup>†</sup>Corresponding Author.



Figure 1. We propose Long Context Tuning (LCT) to expand the context window of pre-trained single-shot video diffusion models. A direct application of LCT is scene-level video generation for short film production, as shown in the top example. We also show several emerging capabilities offered by LCT, such as interactive multi-shot direction and single shot extension, as well as zero-shot compositional generation, despite the model having never been trained on such tasks. *We recommend referring to our Project Page for better visualization.*

trained single-shot video generation model, adapting it to learn a longer context window capable of modeling cross-shot consistency from scene-level video data. To achieve this, we introduce three novel design elements. First, LCT adapts the full attention mechanism from individual shots to encompass all shots within a scene. Inspired by [55], we propose an interleaved 3D Rotary Positional Embedding (RoPE) [47] that assigns unique absolute positions to individual shots while maintaining the relative positional relationships among text-video tokens within each shot. Second, we unify visual condition inputs and diffusion samples by applying independent diffusion timesteps to each shot. This strategy enables joint denoising of all shots, or using

some as conditions by setting their noise to a low level. Finally, we demonstrate the model with bidirectional attention after LCT can be further fine-tuned to context-causal attention, which facilitates efficient auto-regressive generation using KV-cache and substantially reduces computational overhead.

After training on the scene-level video data, experimental results demonstrate that our model exhibits outstanding performance in generating visually and semantically consistent scenes. As shown in Fig. 1, we highlight a generated video with around 20 shots that lasts 3 minutes and keep the appealing visual and semantic consistency. Notably, we find the LCT exhibits emergent capabilities that go beyond

those offered by the pre-trained model it builds upon. For instance, when provided with character identity and environment image (*e.g.*, the bottom of Fig. 1), the model can perform compositional generation by synthesizing videos that seamlessly integrate these elements, despite not being explicitly trained for this task. Moreover, the LCT model also supports autoregressive shot extension, both with and without shortcuts (*e.g.*, the woman walking and car driving examples in Fig. 1 respectively). This feature is particularly useful because it divides long video generations into scene clips, which enables interactive modification by human users. We anticipate that this work could bring inspiration for future research in long video generation.

## 2. Related Works

**Shot-level Video Generation.** Previous literature has extensively explored single-shot video generation. Early approaches [13, 45, 51, 56] relied on GANs but were limited to single-domain datasets. The field later shifted to diffusion-based methods [11, 12, 16, 24, 37, 57, 69, 70, 75], mostly leveraging pre-trained image diffusion models. Among them, representative works such as Align-Your-Latents [6], AnimateDiff [18], Stable Video Diffusion [5], Lumiere [3], and Emu Video [17] extend 2D diffusion UNet with temporal layers to model dynamic motion priors. Later, SoRA [7] demonstrated the scalability of diffusion transformers (DiT) [40] by significantly advancing video generation quality through 3D autoencoding and full-attention mechanisms [52]. This inspired subsequent developments including commercial models like Kling [32], Gen3 [44], and Seaweed, as well as open-source alternatives CogVideoX [63], Mochi [49], and HunyuanVideo [31]. Parallel approaches explore streaming generation [23, 66] and decoder-only language models [30] using discretized visual tokens [67, 68]. Our approach is compatible with single-shot video diffusion models with DiT-based architecture.

**Scene-level Video Generation.** Recent works have addressed scene-level video generation, which requires synthesizing sequential videos depicting continuous events [2, 25, 60, 62]. Research in this field primarily falls into two categories: appearance-conditioned approaches [27, 29] and keyframe-based methods. The former enforces consistency by conditioning on character identity or environment images, while the latter jointly generates initial keyframes before applying independent image-to-video (I2V) [61, 72]. VideoStudio [35] extends text prompts with entity embeddings to preserve appearance information, while MovieDreamer [73] predicts coherent visual tokens that are rendered into I2V keyframes via diffusion decoders. VGoT [74] structures multi-shot generation using identity-preserving embeddings to maintain cross-shot consistency. Other approaches explore feature sharing [1] or re-

trieval augmentation [22, 58]. Nevertheless, existing methods still struggle to model complex scene-level coherence, and keyframe-based approaches are prone to suffer from the ineffective conditioning.

**Long Video Generation.** Current single-shot video models are limited to generating ten-second clips from single prompts. Research addressing this constraint by exploring training-free methods [36, 54], auto-regressive frameworks [66], hierarchical strategy [21, 65], as well as scale up via distributed inference [48]. Among them, FreeNoise [41] reschedules the noise sequence and performs window-based temporal attention. StreamingT2V [23] and CasusVid [66] extend the video length via auto-regressive frame generation. Recent advances incorporate additional control mechanisms [8, 38, 53], including MinT’s [59] time-dependent prompts with specialized position embeddings, as well as DFoT [10, 46]’s history-guided video frame rolling out. Our approach, by contrast, directly enables long video extension via shot extension with and without shortcut.

## 3. Method

Our goal is scene-level video generation that synthesizes multi-shot videos with consistency. To this end, we propose *Long Context Tuning (LCT)* upon the pre-trained single-shot video diffusion model to expand its context window and learn scene-level correlation directly from data.

In Sec. 3.1, we provide preliminaries on single-shot video models. In Sec. 3.2, we detail our approach for LCT. In Sec. 3.3, we explore context-causal attention fine-tuning for efficient auto-regressive generation. In Sec. 3.4, we discuss training and inference implementation.

### 3.1. Preliminary: Single-shot Video Models

We build our method upon a latent video diffusion transformer (DiT) [40] model. The diffusion process operates on the latent representation by encoding RGB video  $x_0$  into  $z_0 = \mathcal{E}(x_0) \in \mathbb{R}^{c' \times h' \times w' \times f'}$ . The model is trained with Rectified Flow (RF) formulation [15, 33, 34], where the noisy sample is a linear interpolation between clean data point and sampled Gaussian noise  $\epsilon$ , *i.e.*,  $z_t = (1-t)z_0 + t\epsilon$ . The training objective is to regress the velocity field, *i.e.*,

$$\mathcal{L} = \mathbb{E}_{t, z_0, \epsilon} \|v_\Theta(z_t, t, c_{text}) - (\epsilon - z_0)\|_2^2, \quad (1)$$

where  $t \in [0, 1]$  is the continuous diffusion timestep,  $v_\Theta(\cdot)$  is the neural network,  $c_{text}$  is the text prompt condition.

The single-shot model employs an MMDiT [15] design, whose transformer block has separate sets of weights for text and video tokens but joins the sequence of the two modalities for self-attention operation. To encode position information, 3D Rotary Position Embedding (RoPE) [47] is applied to the video tokens, where the axis of height, width, and frame index are encoded to different latent channels.

#### [Global Prompt]

**Character 1:** A young man with blonde hair, wearing a dark jacket over a light-colored shirt. He is portrayed as carefree and encouraging; **Character 2:** A young woman with brown hair styled in curls, wearing a dark jacket over a light-colored dress with a necklace; The primary setting is the deck of a ship during the sunset, indicated by the orange and yellow sky. The ship's structure, including railings, masts and superstructure, is visible. **Character 2** encounters **Character 1** on the deck of the ship. **Character 1** suggests **Character 2** trust him and experience something new. He guides her to hold onto the railing, close her eyes, and trust the experience.”

#### [Per-Shot Prompt]



“Close up shot of **Character 1** looking back. The background now shows an orange sky and part of the ship’s mast structure.”

“Medium shot of **Character 2**, standing near a railing. The ship’s structure is in the background. She is speaking.”

“Wide shot showing **Character 2** walks towards **Character 1** near the railing on the right. The orange sky dominates the background.”

“Close up of **Character 1** embracing **Character 2** from behind and instruct her, as she keeps her eyes closed.”

Figure 2. **Scene-level Video Data Example.** *Global prompt* contains shared elements like **character**, **environment**, and **story overview**, while *per-shot prompt* details events in each shot.

## 3.2. Towards Long Context Video Generation

We detail the core components of LCT for scene-level video generation in this section, including data preparation (Sec. 3.2.1) and architecture design (Sec. 3.2.2). We further unify conditioning inputs and diffusion samples via an asynchronous training timesteps strategy (Sec. 3.2.3).

### 3.2.1. Data Curation

**Prompt Structure Definition.** This paper considers scene as a coherent sequence of video clips sharing high-level semantic concepts (characters, environment) while varying in shot-level features (view angles, temporal events). Therefore, we employ a two-tier prompt structure: *global prompts* and *per-shot prompts*, as shown in Fig. 2. Specifically, *global prompts* capture shared elements across shots, following the format “[Character] [Environment] [Story]”, where character descriptions use the structure of “Character [ID]: [Description]”. *Per-shot prompts* detail specific events within each shot and reference characters using “Character [ID]”, rather than common but ambiguous descriptors like “the man/woman”.

**Data Processing.** We collect scene-level video data from public sources across various genres (movies, documentaries, etc.). Raw videos are first segmented into scene videos using scene boundary detection algorithms, then further divided into individual shots via shot-cut detection. We

feed each scene video to Gemini-1.5 [50] and prompt it to provide both global and shot-level descriptions in our specified format. This process yields approximately 500K scene samples averaging 5 shots each. To supplement the training data, we filter some single-shot videos with large temporal variants, divide them into sub segments based on the event changes, and treat them as multi-shot videos. Therefore, adjacent shots transition smoothly without abrupt shotcut, and this process contributes roughly 1M additional samples. We add “[SHOT CUT]” before shot-level prompt in the authentic multi-shot dataset as a special mark.

### 3.2.2. Learning Beyond Single-shot

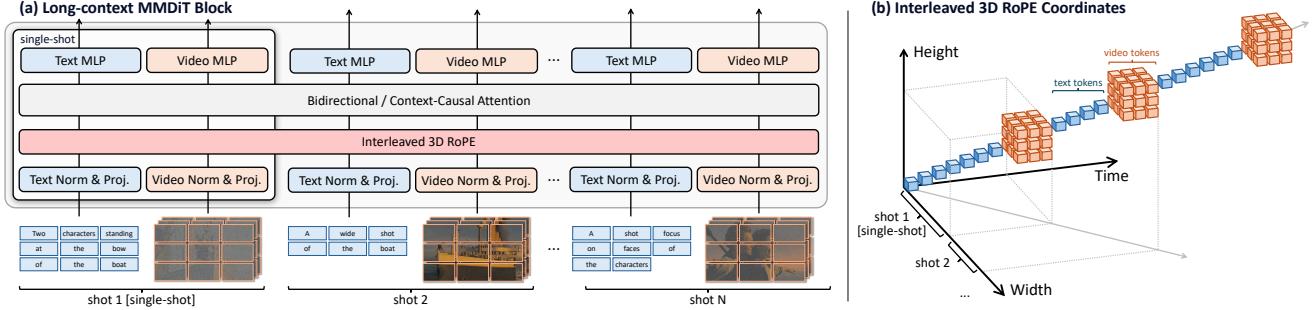
**Long-term Modeling via Full Attention.** Avoiding introducing additional inductive bias, we choose the vanilla full attention [7, 40, 52] to model scene-level consistency. This is done by joining all text and video tokens within the scene, and performing attention operation on the combined sequence jointly, as shown in the Long-context MMDiT in Fig. 3 (a). Note that when the scene only contains one video, the Long-context MMDiT degrades to a single-shot MMDiT. Thus, this framework is still compatible with single shot generation and can be trained with single-shot data to preserve the pre-trained capability.

**Interleaved Position Embedding.** Directly joining all tokens within a scene will hinder the model from identifying which shot the tokens belong to, since tokens from different shots are treated equally. We resolve this issue via an interleaved 3D position embedding, which is conceptually similar to M-RoPE [55] but has not been tested in the diffusion models literature. Specifically, we first add 1D-equivalent 3D RoPE [47] for text-tokens by setting coordinates on three axis identical, and place the video tokens after the end of the text tokens along the space diagonal, as shown in the “shot 1 (single shot)” part in Fig. 3 (b). This is done via efficient single-shot model fine-tuning.

When deal with multiple shots, we keep single-shot’s relative text-video tokens position and append the text-video token group shot by shot, forming an interleaved “[text]–[video]–[text]–...” token sequence along the space diagonal, as shown in Fig. 3 (b). Keeping the relative text-video position allows each shot to inherit the text-visual alignment from the pre-trained model, while different absolute positions distinguish the relation between tokens and the corresponding shot. To deal with global prompt in Sec. 3.2.1, we add dummy video tokens and treat it as a normal text-video pair.

### 3.2.3. Unify Conditions and Diffusion Samples

While the model can generate scene-level multi-shot videos from text prompts with the proposed designs, incorporating visual conditioning enables valuable capabilities like story extension and auto-regressive generation. Unlike existing



**Figure 3. Architecture Designs.** (a) *Long-context MMDiT block*. We expand the attention operation to all text and video tokens within a scene, and apply independent noise levels to individual shots. The interleaved 3D RoPE assigns distinct coordinates for each shot. (b) *Interleaved 3D RoPE coordinates*. At shot-level, text tokens precede video tokens along the space diagonal. At scene-level, tokens are arranged shot by shot, forming an interleaved “[text] – [video] – [text] – . . .” pattern along the space diagonal.

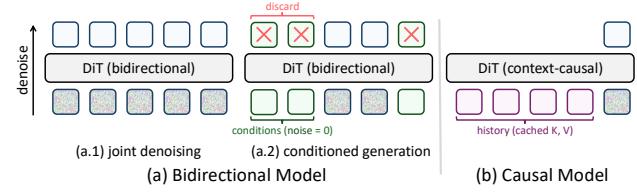
approaches that rely on auxiliary networks for visual conditioning [19, 20, 64, 71], we unify conditioning inputs and diffusion samples through an asynchronous timestep strategy. As shown in Fig. 3 (a), we assign independently sampled diffusion timesteps to each shot during training, rather than applying uniform timesteps across all shots. This establishes dynamic dependencies between shots and encourages the model to exploit cross-shot relationships more effectively. For instance, when a shot within the context window exhibits lower noise compared to others, it naturally serves as a rich source of appearance information to guide the denoising process of noisier shots. Consequently, we can set some samples’ noise to a low level to utilize them as visual conditions, or synchronize diffusion timesteps across all samples for joint generation (Fig. 4 (a)).

### 3.3. Causal Attention Fine-tuning

Our conditioning mechanism enables auto-regressive shot generation by using cleaner history samples as conditions. In this paradigm, information flow is inherently *directional*, *i.e.*, cleaner history samples require little information from subsequent noisy samples, while noisy samples extract cues from preceding history to ensure consistency. This suggests that *bidirectional* attention is redundant and can be replaced with more efficient *causal* attention. After training the bidirectional model with LCT, we therefore implement context-causal attention (bidirectional within each shot, but with tokens attending only to their preceding context) and fine-tune this causal variant. During inference, this architecture allows K, V features cached from history sample generation, eliminating repeated computation and thereby significantly reducing computational overhead (Fig. 4 (b)).

### 3.4. Implementations

**Training.** We jointly train the model on single-shot and scene-level video data to preserve its pre-trained capability. To enable image generation/conditioning, we randomly



**Figure 4. Inference Modes.** (a) *Bidirectional* model enables (a.1) joint or (a.2) visual-conditioned generation, while (b) *context-causal* model supports auto-regressive generation.

substitute shots with one of its single frame at a predetermined probability. Diffusion timesteps are independently sampled from a logit-normal distribution for each shot, with per-shot losses computed according to Eq. (1) and averaged before gradient backpropagation.

**Multi-shot Generation with Human Selection.** Cross-shot dependencies in a scene are often non-sequential, *i.e.*, a shot may refer to elements from distant earlier shots rather than just recent ones. For instance, a character reappearing after several shots requires conditioning on their initial appearance rather than merely on adjacent shots, departing from common auto-regressive generation that is vulnerable to errors accumulation. This insight enables our generation strategy with human selection. We first establish a history pool by generating key character and environmental context shots, then selectively draw from this pool when generating new shots based on relevance rather than recency.

## 4. Experiments

**Training Details.** We implement LCT on a text-to-video diffusion model that adopts MMDiT [15] architecture and is trained on images and videos in their native resolutions and durations [14]. The parameters scale of the pre-trained model is 3B. We use a context window size of nine shots in maximum, and continue the model training with LCT on

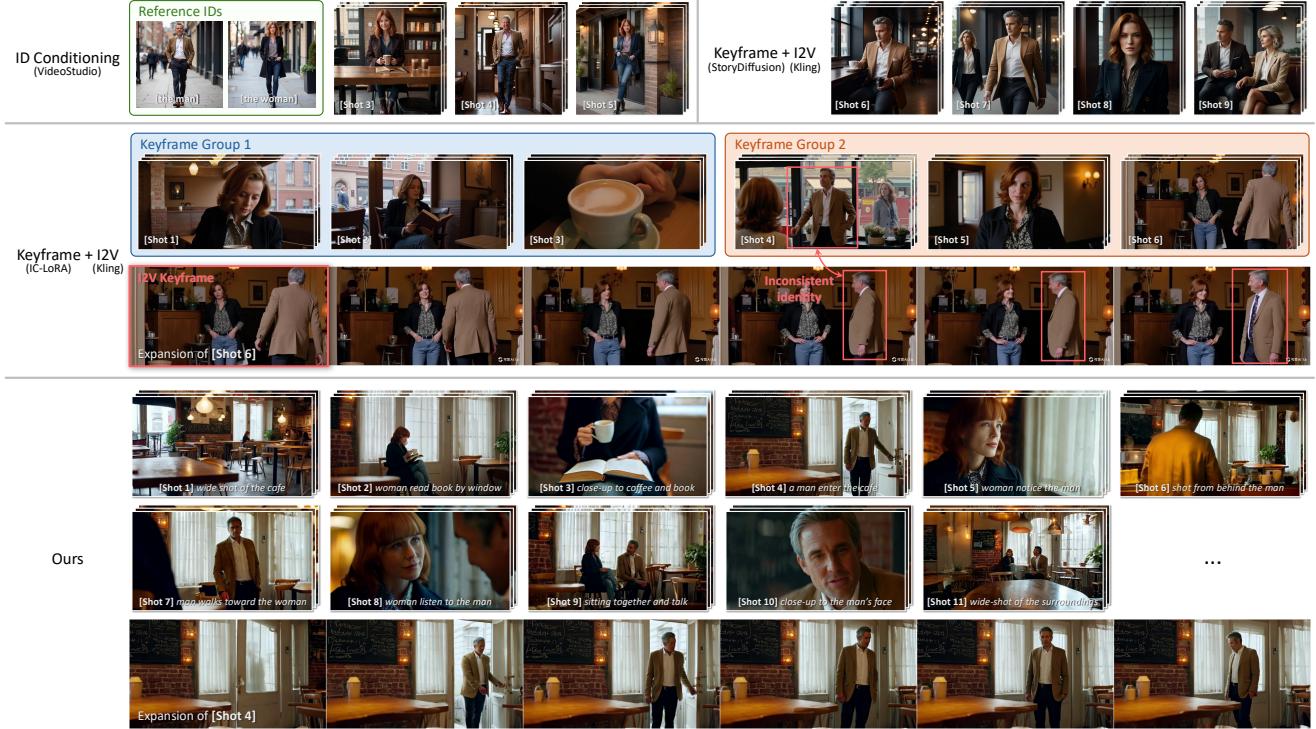


Figure 5. **Qualitative Comparisons.** We show stacked video frames synthesized by all methods, and expand two shots to illustrate the “reappearance” issue discussed in Sec. 4.1. The simplified prompts for each shot can be found in the subtitle in “Ours”.

128 NVIDIA H800s for 135K iterations. For causal attention fine-tuning, we load the model weights after LCT and fine-tune for 9K iterations. The training resolution equals the area size of  $480 \times 480$  with untouched aspect ratio.

#### 4.1. Comparison

**Settings.** We compare our approach against previous arts on scene-level video generation. We focus on two major categories of existing solutions: (1) *appearance-conditioned* approaches, where we adopt VideoStudio [35]; and (2) *keyframe-based* approaches, where we employ two keyframe generation approaches, *i.e.*, FLUX [4] storyboard In-Context LoRA (IC-LoRA) [26] and StoryDiffusion [76], as well as Kling [32] for image-to-video (I2V) animation. Since IC-LoRA only synthesizes three images, we divide a scene into multiple groups and generate keyframes separately. We prompt o3-mini [39] to design the scene and transform the text prompt into method-specific format.

**Qualitative Results.** The qualitative results in Fig. 5 reveal several notable observations. First, baseline methods exhibit limited frame composition diversity. In VideoStudio, character poses closely mirror reference images, while IC-LoRA and StoryDiffusion predominantly generate character-centric mid-shots. Our approach, however, produces diverse framing across wide-shots (*e.g.*, shots 1,

Method	Visual		Temporal		Semantic	
	Aesthetic $\uparrow$	Quality $\uparrow$	Consistency (avg.) $\uparrow$	Text $\uparrow$	User Study $\uparrow$	
VideoStudio [35]	<b>61.68</b>	73.13	95.25	28.00	2.14	
StoryDiffusion [76] + Kling [32]	60.40	<b>74.04</b>	<b>96.57</b>	27.33	<b>2.50</b>	
IC-LoRA [26] + Kling [32]	57.88	69.07	<b>96.27</b>	27.90	1.57	
Ours	<b>60.79</b>	67.44	95.65	<b>30.14</b>	<b>3.79</b>	

Table 1. **Quantitative Evaluations.** We adopt automatic metrics and average human ranking (AHR). “Consistency (avg.)” represents the average score of subject and background consistency.

11), mid-shots (*e.g.*, shots 2, 4, 6), and close-up shots (*e.g.*, shots 3, 8, 10), enabling richer narrative perspectives. Second, baselines demonstrate deficient prompt alignment. IC-LoRA notably fails to generate an establishing shot (*i.e.*, environment without characters) for shot 1, instead introducing extraneous characters in other shots. This stems from concatenating shot descriptions into a single prompt, impeding the model to disentangle semantic elements. VideoStudio and StoryDiffusion tend to synthesize characters’ front-facing views and thus compromise prompt alignment. Third, our method achieves superior consistency. The female character maintains her position beside the window throughout the scene, whereas baselines show disruptive positional shifts that break narrative continuity.

Finally, we highlight an often overlooked “reappearance” issue inherent in keyframe-based scene-level video generation approaches. As evident in the “Expansion of

Task	Model	Settings	Aesthetic $\uparrow$	Quality $\uparrow$	Consistency (avg.) $\uparrow$	Text $\uparrow$
single-shot	pre-trained	-	52.09	63.37	94.98	30.81
	post-LCT	-	55.91	64.23	95.06	32.22
multi-shot	bidirectional	joint denoise	56.14	57.68	95.34	29.33
	bidirectional	auto-regressive	55.46	56.62	94.87	30.24
	context-causal	auto-regressive	55.49	50.31	94.96	30.15

Table 2. **Quatitative Ablation Studies.** We compare single-shot generation capabilities in the upper part, and the effects of inference modes in the lower part.

[Shot 6]” in the “Keyframe (IC-LoRA) + I2V (Kling)” approach, when a character’s identity is not fully captured in the initial keyframe but appears in subsequent I2V video frames, its consistency across shots cannot be guaranteed. Conversely, our direct video generation solution effectively eliminates this problem, as demonstrated in the “Expansion of [Shot 4]” in our results.

**Quantitative Results.** We evaluate all methods across three dimensions: visual quality, temporal consistency, and semantic coherence. For automatic assessment, we utilize VBVench [28] that focuses on shot-level quality. Given the absence of scene-level coherence metrics, we employ user study and instruct participants to prioritize cross-shot consistency over appearance. As shown in Tab. 1, while our method performs slightly below visually-conditioned baselines in visual quality, it substantially outperforms them in semantic alignment, demonstrating superior capability in maintaining scene-level coherence.

## 4.2. Ablative Studies

In this section, we discuss the model’s training and inference behaviors of Long Context Tuning (LCT).

**Single-shot Generation after LCT.** Our proposed LCT introducing *no additional parameters*, enlarging the model capability under the same capacity. Therefore, it is important to evaluate whether LCT sacrifices the pre-trained single-shot generation ability. We investigate this by synthesizing videos using an established prompt suite with models before and after LCT, and evaluate them with VBVench [28]. As shown in the upper part of Tab. 2, the post-LCT model demonstrates consistent performance gains in all metrics, indicating that LCT not only maintains but potentially enhances single-shot generation capability.

**Effects of Inference Modes.** Our approach supports multiple inference modes: joint generation and auto-regressive (conditional) generation for bidirectional model (Fig. 4 (a)), as well as auto-regressive generation for the context-causal model (Fig. 4 (b)). To assess how they affect generation quality, we evaluate videos produced by each mode. The lower section of Tab. 2 reveals the improved text alignment from joint generation to auto-regressive settings. This likely occurs because in auto-regressive modes, later shots can access preceding history



Figure 6. **Fidelity to History Condition.** The video background generated by the causal model exhibits superior fidelity to the history condition, as evidenced by the street lights’ layout.



Figure 7. **Effects of Conditioning Timestep.** Large timesteps sacrifice fidelity to the condition.

as complementary information to the text prompts.

Moreover, we found that the context-causal architecture demonstrates superior fidelity to history conditions compared to the bidirectional one, as evident in Fig. 6. We attribute this to the causal attention mechanism, which enforces sequential dependency on history and consequently assigns greater weight to preceding conditions.

**History Conditioning Timestep.** In auto-regressive inference, previously generated samples are perturbed to noise level  $t_c$  and utilized as conditioning inputs, as shown in Fig. 4 (a.2) and (b). Intuitively, using clean histories ( $t_c = 0$ ) preserves maximum information. However, this leads to the error accumulation issue, with generated artifacts propagated and amplified through subsequent samples. We therefore examined the effect of conditioning timestep in the auto-regressive setting. As shown in Fig. 7, substantial history details are preserved even with a relatively high timestep such as  $t_c = 500$ , while at  $t_c = 900$ , the overall structure remains though high-frequency details are lost. Further analysis in Fig. 8 demonstrates that low conditioning timesteps result in rapid quality degradation during sequential shot generation, while higher timesteps mitigate this problem at the cost to history fidelity. Based on these findings, we employ  $t_c = 100 \sim 500$  in practice to optimize the balance between long-term generation quality and fidelity to historical context.

**Causal Fine-tuning Adaptation.** We evaluate the efficiency of adapting a bidirectional model post-LCT to a context-causal architecture. In Fig. 9, we track performance evolution during adaptation by evaluating multiple checkpoints under identical conditions, with particular emphasis on identity consistency. As illustrated, when initially loading bidirectional weights into the context-causal architecture (step 0), the model remains uncollapsed but generates frames with poor identity consistency. After only 1K up-



Figure 8. **Effects of History Timestep.** Large timesteps mitigate “error accumulation” issue at the cost of history fidelity.

dates, identity consistency improves dramatically, achieving quality comparable to the final 9K-step checkpoint. We attribute this efficiency to the bidirectional model’s conditioning mechanism: the preceding clean history conditions rarely need to attend to subsequent frames since they primarily contain their own information. This behavior inherently resembles the unidirectional information flow enforced by the context-causal architecture, facilitating rapid adaptation between the two paradigms.

### 4.3. Emerging Capabilities

In this section, we discuss several emerging model capabilities after Long Context Tuning (LCT).

**Conditional and Compositional Generation.** Our model inherently supports diverse conditioning with images and videos. As demonstrated in the second example of Fig. 1, we can generate narrative continuations by using an existing video as the initial shot condition. Remarkably, despite no explicit training for this capability, our model enables compositional generation by accepting separate identity and environment images to synthesize coherent videos integrating these distinct elements, as illustrated in the final example of Fig. 1. This ability emerges from the model’s learned scene-level visual relations in the training corpus, where scenes frequently contain establishing environmental shots, character close-ups, and integrated shots depicting character-environment interactions. Our findings offer a fresh perspective on appearance-conditioned video synthesis that eliminates task-specific data curation, providing a more generalizable approach to compositional generation.

**Single-shot Extension.** Benefit from training on single-shot segments (Sec. 3.2.1), our model enables interactive shot extension without shotcut. For this application, we remove the “[SHOT CUT]” and craft bridging prompts that seamlessly connect existing content with desired future content. As shown in the third example in Fig. 1, this approach



Figure 9. **Causal Adaptation.** After 1K updates from bidirectional weights, the causal architecture shows excellent consistency.

can extend a single shot to minute-long durations, auto-regressively generating 10-second segments while maintaining visual consistency.

**Interactive Generation.** Our model enables interactive multi-shot development, as shown in the second example of Fig. 1. This facilitates an iterative workflow where directors can progressively shape content shot-by-shot based on previously generated footages, eliminating the need for comprehensive upfront prompting and allowing for creative decision-making with immediate visual feedback.

## 5. Conclusion

We propose *Long Context Tuning (LCT)* to adapt single-shot video models for scene-level generation. By expanding context window to the entire scene, implementing interleaved 3D position embeddings, and employing asynchronous training timesteps, we enable flexible scene synthesis without additional parameters. Our approach also supports context-causal attention for efficient auto-regressive generation with KV-cache. Evaluations show our method outperforms existing approaches in scene-level video generation while exhibiting emerging capabilities like compositional generation and shot extension. We anticipate it will reveal new possibilities for video generation.

**Discussion.** As LCT allows users to create shots following their intents, we believe that in the future, leveraging the power of Multimodal Large Language Models (MLLMs) as planners for scene-level video generation would be a great alternative for further applications (*e.g.*, story generation, even reasoning from generative perspectives). Besides, current auto-regressive generation includes all preceding tokens as history conditions, which is effective yet introduces potential redundancy. Therefore, involving token dynamic routing into attention mechanisms may be a promising and practical direction for future work.

## References

- [1] Yuval Atzmon, Rinon Gal, Yoad Tewel, Yoni Kasten, and Gal Chechik. Multi-shot character consistency for text-to-video generation. *arXiv preprint arXiv:2412.07750*, 2024. 3
- [2] Hritik Bansal, Yonatan Bitton, Michal Yarom, Idan Szpektor, Aditya Grover, and Kai-Wei Chang. Talc: Time-aligned captions for multi-scene text-to-video generation. *arXiv preprint arXiv:2405.04682*, 2024. 3
- [3] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Hermann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3
- [4] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 6
- [5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1, 3
- [6] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3
- [7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. 1, 3, 4
- [8] Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. Dictctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. *arXiv preprint arXiv:2412.18597*, 2024. 3
- [9] Vasileios T Chasanis, Aristidis C Likas, and Nikolaos P Galatsanos. Scene detection in videos using shot clustering and sequence alignment. *IEEE transactions on multimedia*, 11(1):89–100, 2008. 1
- [10] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2025. 3
- [11] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 3
- [12] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 3
- [13] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. *ACM Transactions on Graphics (TOG)*, 39(4):75–1, 2020. 3
- [14] 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. 5
- [15] 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. 3, 5
- [16] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 3
- [17] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Factorizing text-to-video generation by explicit image conditioning. In *European Conference on Computer Vision*, pages 205–224. Springer, 2024. 3
- [18] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3
- [19] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348. Springer, 2024. 5
- [20] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 5
- [21] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 3
- [22] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*, 2023. 3
- [23] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024. 3
- [24] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video dif-

- fusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 3
- [25] Panwen Hu, Jin Jiang, Jianqi Chen, Mingfei Han, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. Storyagent: Customized storytelling video generation via multi-agent collaboration. *arXiv preprint arXiv:2411.04925*, 2024. 3
- [26] 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, 6
- [27] Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*, 2025. 1, 3
- [28] Ziqi Huang, Yinan He, Jiahuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 7
- [29] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6689–6700, 2024. 1, 3
- [30] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vignesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 3
- [31] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanyvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 3
- [32] Kuaishou. Kling video model. <https://kling.kuaishou.com/en>, 2024. 1, 3, 6
- [33] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3
- [34] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3
- [35] 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, 3, 6
- [36] Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. *Advances in Neural Information Processing Systems*, 37:131434–131455, 2025. 3
- [37] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023. 3
- [38] Gyeongrok Oh, Jaehwan Jeong, Sieun Kim, Wonmin Byeon, Jinkyu Kim, Sungwoong Kim, and Sangpil Kim. Mevg: Multi-event video generation with text-to-video models. In *European Conference on Computer Vision*, pages 401–418. Springer, 2024. 3
- [39] OpenAI. o3-mini. <https://openai.com/index/openai-o3-mini/>, 2024. 6
- [40] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1, 3, 4
- [41] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023. 3
- [42] Anyi Rao, Lining Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10155, 2020. 1
- [43] Zeeshan Rasheed and Mubarak Shah. Scene detection in hollywood movies and tv shows. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, pages II–343. IEEE, 2003. 1
- [44] RunwayML. Gen-3 alpha. <https://runwayml.com/research/introducing-gen-3-alpha>, 2024. 1, 3
- [45] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhosseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3626–3636, 2022. 3
- [46] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025. 3
- [47] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 2, 3, 4
- [48] Zhenxiong Tan, Xingyi Yang, Songhua Liu, and Xinchao Wang. Video-infinity: Distributed long video generation. *arXiv preprint arXiv:2406.16260*, 2024. 3
- [49] Genmo Team. Mochi 1. <https://github.com/genmoai/models>, 2024. 3
- [50] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 4
- [51] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018. 3
- [52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4

- [53] Ruben Villegas, Mohammad Babaizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 3
- [54] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023. 3
- [55] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 4
- [56] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1160–1169, 2020. 3
- [57] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pages 1–20, 2024. 3
- [58] Zun Wang, Jialu Li, Han Lin, Jaehong Yoon, and Mohit Bansal. Dreamrunner: Fine-grained storytelling video generation with retrieval-augmented motion adaptation. *arXiv preprint arXiv:2411.16657*, 2024. 3
- [59] Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varnith Chordia, Igor Gilitschenski, and Sergey Tulyakov. Mind the time: Temporally-controlled multi-event video generation. *arXiv preprint arXiv:2412.05263*, 2024. 3
- [60] Zhifei Xie, Daniel Tang, Dingwei Tan, Jacques Klein, Tegawend F Bissyand, and Saad Ezzini. Dreamfactory: Pioneering multi-scene long video generation with a multi-agent framework. *arXiv preprint arXiv:2408.11788*, 2024. 3
- [61] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024. 1, 3
- [62] Dingyi Yang, Chunru Zhan, Ziheng Wang, Biao Wang, Tiezheng Ge, Bo Zheng, and Qin Jin. Synchronized video storytelling: Generating video narrations with structured storyline. *arXiv preprint arXiv:2405.14040*, 2024. 3
- [63] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazhen Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3
- [64] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ipadapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 5
- [65] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023. 3
- [66] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast causal video generators. *arXiv preprint arXiv:2412.07772*, 2024. 3
- [67] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469, 2023. 3
- [68] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Verhari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 3
- [69] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8850–8860, 2024. 3
- [70] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, pages 1–15, 2024. 3
- [71] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 5
- [72] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 3
- [73] Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequence. *arXiv preprint arXiv:2407.16655*, 2024. 1, 3
- [74] Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, et al. Videogen-of-thought: A collaborative framework for multi-shot video generation. *arXiv preprint arXiv:2412.02259*, 2024. 1, 3
- [75] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 3
- [76] 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, 6