ShotStream: Streaming Multi-Shot Video Generation for Interactive Storytelling

Yawen Luo[1⋆], Xiaoyu Shi[2], Junhao Zhuang[1], Yutian Chen[1], Quande Liu[2], Xintao Wang[2], Pengfei Wan[2], and Tianfan Xue[1,3,✉]

[1]MMLab, CUHK [2]Kuaishou Technology
[3]CPII under InnoHK ✉ Corresponding author

Fig. 1: Multi-Shot Video results of ShotStream. ShotStream is an autoregressive multi-shot video generation model enabling interactive storytelling and on-the-fly synthesis at 16 FPS on a single GPU. Each case presented here (rows 1–4) illustrates a generated sequence comprising five consecutive shots and 405 total frames, demonstrating the model's capacity to maintain narrative and visual consistency across scene transitions. The expanded bottom row details the high visual quality achieved within a single shot. We highly encourage readers to view our supplementary materials for the video results.

Abstract. Multi-shot video generation is crucial for long narrative storytelling, yet current bidirectional architectures suffer from limited interactivity and high latency. We propose ShotStream, a novel causal multi-shot architecture that enables interactive storytelling and efficient on-the-fly frame generation. By reformulating the task as next-shot generation conditioned on historical context, ShotStream allows users to dynamically instruct ongoing narratives via streaming prompts. We achieve

⋆ Work done during an internship at Kling team, Kuaishou Technology.

this by first fine-tuning a text-to-video model into a bidirectional next-shot generator, which is then distilled into a causal student via Distribution Matching Distillation. To overcome the challenges of inter-shot consistency and error accumulation inherent in autoregressive generation, we introduce two key innovations. First, a *dual-cache memory mechanism* preserves visual coherence: a *global context cache* retains conditional frames for inter-shot consistency, while a *local context cache* holds generated frames within the current shot for intra-shot consistency. And a RoPE discontinuity indicator is employed to explicitly distinguish the two caches to eliminate ambiguity. Second, to mitigate error accumulation, we propose a two-stage distillation strategy. This begins with intra-shot self-forcing conditioned on ground-truth historical shots and progressively extends to inter-shot self-forcing using self-generated histories, effectively bridging the train-test gap. Extensive experiments demonstrate that ShotStream generates coherent multi-shot videos with sub-second latency, achieving 16 FPS on a single GPU. It matches or exceeds the quality of slower bidirectional models, paving the way for real-time interactive storytelling. Code and models will be released.

Keywords: Multi-shot Long Video Generation · Efficient Causal Architecture · Interactive Storytelling

1 Introduction

While current text-to-video models [2,18,28,33] excel at synthesizing high-fidelity single-shot videos, the field is rapidly advancing toward long-form narrative storytelling [10] akin to traditional film and television. This evolution necessitates multi-shot video generation, which enables the creation of sequential shots that maintain subject and scene consistency while advancing the narrative through varied content. For instance, cinematic techniques like the *shot-reverse shot* [11] create cohesive interactions by cutting back and forth between characters, effectively guiding the viewer's attention through dynamic perspectives. Driven by the growing demand for such complex cinematic narratives, multi-shot video generation [1, 4, 26, 34, 41, 49] has gained increased attention.

Existing multi-shot video generation methods [4,10,14,23,26,34,36,41] mainly rely on bidirectional architectures to model intra-shot and inter-shot dependencies, ensuring temporal and narrative consistency. Although effective, these bidirectional architectures suffer from two main limitations: 1) *Lack of interactivity*: Current methods require all prompts upfront to generate the entire multi-shot sequence at once, making it difficult to adjust individual shots without a complete re-generation. A more user-friendly approach would accept streaming prompt inputs at runtime, enabling users to interactively guide the narrative and adapt the current shot based on previously generated content. 2) *High latency*: The computational cost of bidirectional attention grows quadratically with context length, posing a major challenge for long sequences. Even with the integration of sparse attention mechanisms to reduce overhead and accelerate generation,

	Multi-Shot	Long Gen.	Interactive	Real-Time
Bidirectioal Multi-Shot Models	✓	✗	✗	✗
Long Autoregressive Models	✗	✓	✓	✓
ShotStream (Ours)	✓	✓	✓	✓

Fig. 2: Overview of the ShotStream workflow, which enables real-time, long, multi-shot video generation from streaming prompts.

these models still exhibit prohibitive latency. For instance, HoloCine [23] requires approximately 25 minutes to generate a 240-frame multi-shot video.

To overcome these limitations, we propose ShotStream, a novel causal multi-shot architecture that enables interactive storytelling and efficient on-the-fly frame generation. To achieve interactivity, we reformulate multi-shot synthesis as an autoregressive next-shot generation task, where each subsequent shot is generated by conditioning on previous shots. This reformulation allows ShotStream to accept streaming prompts as inputs and generate videos shot-by-shot, empowering users to dynamically guide the narrative at runtime by adjusting content, altering visual styles, or introducing new characters, as shown in Fig. 2.

To achieve this efficient causal architecture, we first train a bidirectional teacher model for next-shot prediction, conditioned on historical context. Because past shots comprise hundreds of frames, retaining the entire history introduces severe temporal redundancy and becomes memory-prohibitive. To address this, we condition the model on a sparse subset of historical frames rather than the entire sequence. Specifically, we introduce a dynamic sampling strategy that selects frames based on the number of preceding shots and specific conditional constraints, effectively preserving historical information within a strict frame budget. We then inject these sampled context frames by concatenating their context tokens with noise tokens along the temporal dimension to form a unified input sequence. This concatenation-based injection mechanism is highly parameter-efficient and eliminates the need for additional architectural modules.

Subsequently, we distill this slow, multi-step bidirectional teacher model into an efficient, 4-step causal student model via Distribution Matching Distillation [46, 47]. However, transitioning to this causal architecture introduces two primary challenges: 1) maintaining consistency across shots, and 2) preventing error accumulation to sustain visual quality during autoregressive generation. To address the first challenge, we introduce a novel *dual-cache memory mechanism.*

A *global context cache* stores sparse conditional historical frames to ensure inter-shot consistency, while a *local context cache* retains frames generated within the current shot to preserve intra-shot continuity. Naively combining these caches introduces ambiguity, as the causal model struggles to differentiate between historical and current-shot contexts. To resolve this, we propose a RoPE discontinuity indicator that explicitly distinguishes between the global and local caches.

The second challenge, error accumulation, stems primarily from the train-test gap [12]. We mitigate this challenge by aligning training with inference through a proposed two-stage progressive distillation strategy. We begin with intra-shot self-forcing conditioned on ground-truth historical shots, where the generator rolls out the current shot causally, chunk-by-chunk, to establish foundational next-shot capabilities. We then progressively transition to inter-shot self-forcing using self-generated histories. In this stage, the model rolls out the multi-shot video shot-by-shot, while generating the internal frames of each individual shot chunk-by-chunk. This strategy bridges the train-test gap and significantly enhances the quality of autoregressive multi-shot generation.

Extensive evaluations demonstrate that ShotStream generates long, narratively coherent multi-shot videos (as shown in Fig. 1) while achieving an efficient 16 FPS on a single NVIDIA H200 GPU. Quantitatively, our method achieves state-of-the-art performance on the test set regarding visual consistency, prompt adherence, and shot transition control. To complement these metrics with subjective evaluation, we conduct a user study involving 54 participants. Users are asked to compare 24 multi-shot videos generated by our method against those from baselines. The results reveal a decisive user preference for ShotStream in terms of visual consistency, overall visual quality, and prompt adherence.

In summary, our main contributions are as follows:

- We present ShotStream, *a novel causal multi-shot long video generation architecture* that enables interactive storytelling and on-the-fly synthesis.
- We reformulate multi-shot synthesis as a next-shot generation task to support interactivity, allowing users to dynamically adjust ongoing narratives via streaming prompts.
- We design a novel dual-cache memory mechanism for our causal model to ensure both inter-shot and intra-shot consistency, coupled with a RoPE discontinuity indicator to explicitly distinguish between the two caches.
- We propose a two-stage distillation strategy that effectively mitigates error accumulation by bridging the gap between training and inference to enable robust, long-horizon multi-shot generation.

2 Related Work

2.1 Multi-Shot Video Generation

Driven by interest in narrative video generation, multi-shot video synthesis has advanced rapidly [4, 10, 14, 23, 26, 34, 36, 41, 51]. Current methods generally fall into two categories. Keyframe-based approaches [41, 50, 51] generate the initial

frames of each shot and extend them using image-to-video models. However, they often struggle with global coherence, as consistency is enforced only at the keyframe level while intra-shot content remains isolated. The second paradigm, unified sequence modeling [4,10,14,23,26,34], jointly processes all shots within a sequence. For instance, LCT [10] applies full attention across all shots using interleaved 3D position embeddings to distinguish them. While efficiency-focused variants like MoC [4] and HoloCine [23] employ dynamic or sparse attention patterns to reduce computational burden, they still suffer from high latency. Furthermore, their bidirectional architectures and unified modeling inherently limit interactivity, complicating the adjustment of specific shots within a generated sequence.

2.2 Autoregressive Long Video Generation

Driven by next-token prediction objectives, autoregressive models naturally support the gradual rollout required for long video generation [3, 38, 39]. Recently, integrating autoregressive modeling with diffusion has emerged as a promising paradigm for causal, high-quality video synthesis [6,9,12,20,21,32,42,44,45,48]. Methods like CausVid [48] achieve low-latency streaming by distilling multi-step diffusion into a 4-step causal generator. To mitigate exposure bias from the train-test sequence length discrepancy, Self Forcing [12] and Rolling Forcing [20] condition generation on self-generated outputs and progressive noise levels, respectively, to suppress error accumulation. Additionally, LongLive [42] enables dynamic runtime prompting via a KV-recache mechanism. Despite these advancements, existing techniques are largely confined to single-scene generation and struggle with multi-shot narratives. Our method addresses this gap, extending autoregressive modeling to generate cohesive, multi-shot narrative videos.

3 Preliminary

3.1 Distribution matching distillation

Distribution Matching Distillation (DMD) [46,47] distills slow, multi-step diffusion models into fast, few-step student generators while maintaining high quality. The key objective is to match the student and teacher at the distribution level by minimizing the reverse KL divergence between the smoothed data distribution, $\boldsymbol{p}_{\text{data}}$, and the student generator's output distribution, $\boldsymbol{p}_{\text{gen}}$. This optimization is performed across random timesteps t, where the gradient is approximated by the difference between two score functions: one trained on the true data distribution and another trained on the student generator's output distribution using a denoising loss. Further details are provided in the supplementary materials.

3.2 Self Forcing

Error accumulation [24,29] remains a persistent challenge in autoregressive video generation, caused by the discrepancy between using ground-truth data during

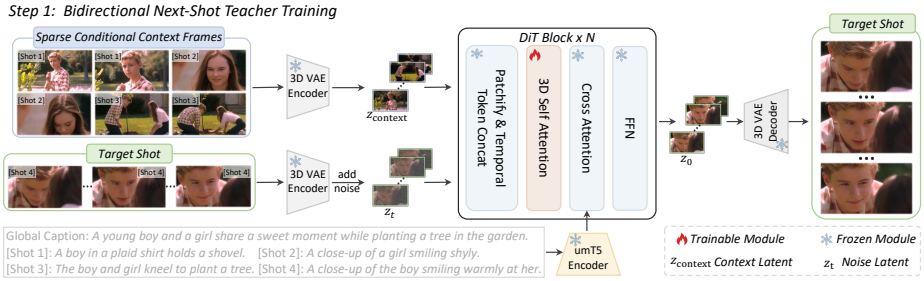

Fig. 3: Architecture of the Bidirectional Next-Shot Teacher Model. To realize ShotStream, we first fine-tune a text-to-video model into a bidirectional next-shot model, which generates subsequent shots conditioned on sparse context frames from preceding shots. These conditional context frames are encoded into latents via a 3D VAE and injected by concatenating them with noise latents along the temporal dimension. Notably, only the 3D spatial-temporal attention layers within the DiT Blocks are optimized during fine-tuning. A 4-shot example is shown here for illustration.

training and relying on imperfect predictions during inference. To bridge this train-test gap, self forcing [12] introduces a training paradigm that explicitly unrolls the autoregressive process. By conditioning each frame on previously generated outputs rather than ground-truth frames, the model is compelled to navigate and recover from its own inaccuracies. Consequently, self forcing [12] effectively mitigates exposure bias and stabilizes long generation.

4 Method

This section details the architecture and training methodology of ShotStream. We first fine-tune a text-to-video model into a bidirectional next-shot model (Sec. 4.1). This model is subsequently distilled into an efficient, 4-step causal model via Distribution Matching Distillation. We also propose a novel dual-cache memory mechanism and a two-stage distillation strategy to enable efficient, robust, and long-horizon multi-shot generation (Sec. 4.2).

4.1 Bidirectional Next-Shot Teacher Model

The objective of the next-shot model is to generate a subsequent shot conditioned on historical shots. Since historical shots contain hundreds of frames with high visual redundancy, retaining the entire history is unnecessary and impractical with a limited conditional budget. Therefore, we condition the model on sparse context frames extracted via a dynamic sampling strategy. Specifically, given S_{hist} historical shots and a maximum conditional context budget of f_{context} frames, we sample $\lfloor f_{\text{context}}/S_{\text{hist}} \rfloor$ frames from each historical shot, where $\lfloor \cdot \rfloor$ denotes the floor function. Any remaining budget is allocated to the most recent shot to fully utilize the budget, which is set to 6 frames in our experiments.

To condition the model on sampled sparse context frames V_{context}, we employ a temporal token concatenation mechanism, an injection technique proven effective across multi-control generation [15], editing [43], and camera motion cloning [19, 22]. Although effective, these methods do not distinguish between the captions of condition frames and target frames; instead, they uniformly apply the target frame's caption to the condition frames. Directly adopting this approach for next-shot generation is problematic, as the captions of previous shots contain crucial information that binds past visual information to textual descriptions. This binding facilitates the extraction of necessary context for generating the subsequent shot. Therefore, we also inject the specific captions corresponding to each conditional context frame into the model, *i.e.*, the frames of each shot attend to both the global caption and the corresponding local shot caption via cross-attention. Specifically, as shown in Fig. 3, our next-shot model reuses the 3D VAE ε from the base model to transform V_{context} into conditioning latents,

$$z_{\text{context}} = \varepsilon(V_{\text{context}})\,, \tag{1}$$

where $z_{\text{context}} \in \mathbb{R}^{f_{\text{context}} \times c \times h \times w}$ comprises f_{context} frames, c channels, and a spatial resolution of $h \times w$. Building upon this shared latent space, we first patchify the condition latent z_{context} and the noisy target latent $z_t \in \mathbb{R}^{f \times c \times h \times w}$ with f frames into tokens:

$$x_j = \text{Patchify}(z_j), z_j \in \{z_{\text{context}}, z_t\}\,. \tag{2}$$

The resulting condition tokens x_{context} and noisy video tokens x_t are then concatenated along the frame dimension to form the input for the DiT blocks:

$$x_{\text{input}} = \text{FrameConcat}(x_{\text{context}}, x_t)\,. \tag{3}$$

The notation FrameConcat denotes that the condition tokens are concatenated with the noise tokens along the frame dimension. Given that the token sequences $x_{\text{context}} \in \mathbb{R}^{b \times f_{\text{context}} \times s \times d}$ and $x_{\text{t}} \in \mathbb{R}^{b \times f \times s \times d}$ share the same batch size b, spatial token count s of each frame, and feature dimension d, this temporal concatenation yields a combined tensor $x_{\text{input}} \in \mathbb{R}^{b \times (f_{\text{context}}+f) \times s \times d}$. During the training process, noise is added exclusively to the target video tokens, keeping the context tokens clean. This design enables the DiT's native 3D self-attention layers to directly model interactions between the condition and noise tokens, without introducing new layers or parameters to the base model.

4.2 Causal Architecture and Distillation

The bidirectional next-shot teacher model (detailed in Sec. 4.1) requires approximately 50 denoising steps, resulting in high inference latency. To enable low-latency generation, we distill this multi-step teacher into an efficient 4-step causal generator. However, transitioning to this causal architecture introduces two primary challenges: 1) maintaining consistency across shots, and 2) preventing error accumulation to sustain visual quality during autoregressive generation.

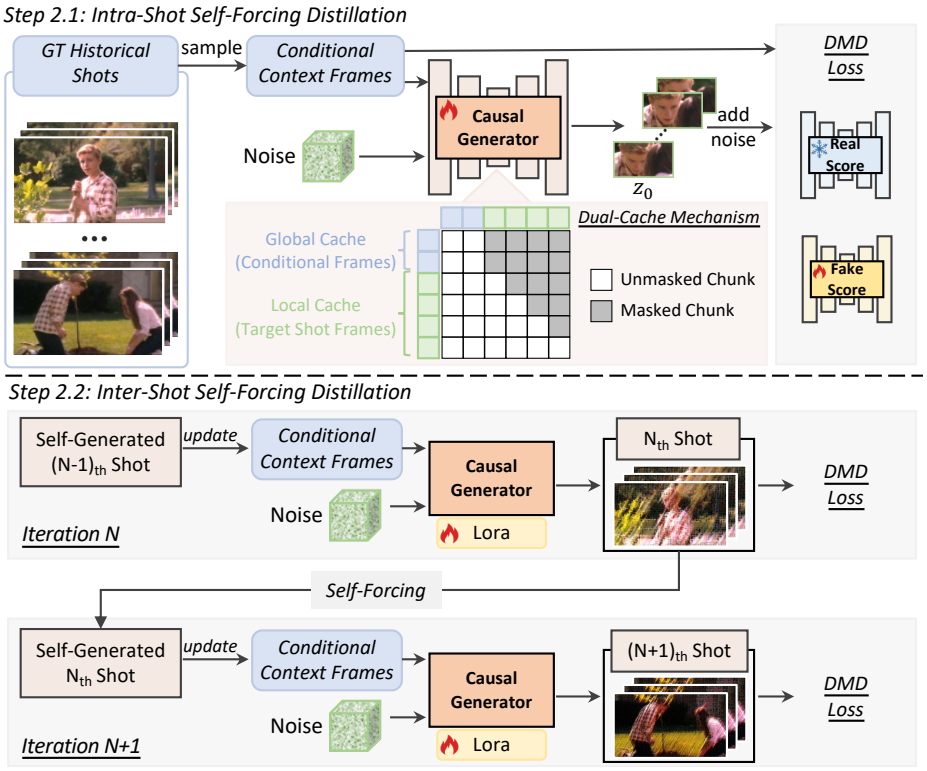

Fig. 4: Causal Architecture and Two-Stage Distillation Pipeline. We distill a slow, multi-step bidirectional teacher into an efficient, few-step causal generator. To maintain visual coherence, we propose a novel dual-cache memory mechanism: a global context cache stores conditional frames to ensure inter-shot consistency, while a local context cache retains generated frames within the target shot to guarantee intra-shot consistency. To prevent error accumulation, we employ a progressive two-stage distillation strategy. In the first stage, intra-shot self-forcing distillation (Step 2.1), the model is conditioned on ground-truth historical shots to causally generate the current shot chunk-by-chunk. In the second stage, inter-shot self-forcing distillation (Step 2.2), the model is conditioned on its own previously generated shots, rolling out the video shot-by-shot while iteratively generating the frames of each individual shot chunk-by-chunk.

To address these issues, we propose two key innovations: a dual-cache memory mechanism and a two-stage distillation strategy, respectively.

Dual-Cache Memory Mechanism. To maintain visual coherence, we introduce a novel dual-cache memory mechanism (Fig. 4): a global cache stores sparse conditional frames to preserve inter-shot consistency, while a local cache retains recently generated frames to ensure intra-shot consistency. However, querying both caches simultaneously within our chunk-wise causal architecture introduces temporal ambiguity, as the model struggles to distinguish historical from current-

shot contexts. To address this, we propose a discontinuous RoPE strategy that explicitly decouples the global and local contexts by introducing a discrete temporal jump at each shot boundary. Specifically, for the t-th latent z_t within the k-th shot, its temporal rotation angle is formulated as $\Theta_t = \phi t + k\theta$, where ϕ denotes the base temporal frequency and θ serves as the phase shift representing the shot-boundary discontinuity.

Two-stage Distillation Strategy. A major challenge in autoregressive multi-shot video generation is error accumulation caused by the training-inference gap [12]. To mitigate this, we propose a two-stage distillation training strategy. In the first stage, **intra-shot self-forcing** (Fig. 4, Step 2.1), the model samples global context frames from ground-truth historical shots while the chunk-wise causal generator produces the target shot via a temporal autoregressive rollout. Specifically, the local cache utilizes previously self-generated chunks from the current target shot rather than ground-truth data. Although this stage establishes foundational next-shot generation capabilities, a training-inference gap remains: during inference, the model must condition on its own potentially imperfect historical shots instead of the ground truth. To bridge this gap, we introduce the second stage: **inter-shot self-forcing** (Fig. 4, Step 2.2). Specifically, the causal model generates the initial shot from scratch and applies DMD. For all subsequent iterations, the generator synthesizes the next shot conditioned entirely on prior self-generated shots. During each iteration, the model continues to employ intra-shot self-forcing to generate each new shot chunk by chunk, applying DMD exclusively to the newly generated shot. This autoregressive unrolling continues until the entire multi-shot video is generated. By closely mirroring the inference-time rollout, this stage aligns training and inference, effectively mitigating error accumulation and enhancing overall visual quality.

Inference. The inference procedure of ShotStream identically aligns with its training process. ShotStream generates multi-shot videos in a shot-by-shot manner. As each new shot is generated, the global context frames are updated by sampling from previously synthesized historical shots. Within the current shot, video frames are generated sequentially chunk by chunk, leveraging our causal few-step generator and KV caching to ensure computational efficiency.

5 Experiments

5.1 Experiment Setup

Implement Details. We build ShotStream upon Wan2.1-T2V-1.3B [33] to generate 832×480 video clips. The bidirectional next-shot teacher is trained on an internal dataset of 320K multi-shot videos. For causal adaptation, the student model is initialized via regression on 5K teacher-sampled ODE solution pairs [48]. Distillation proceeds in two stages: intra-shot self-forcing using ground-truth historical shots from the dataset, followed by inter-shot self-forcing using captions from a subset of 5-shot videos. Architecturally, the model operates with a chunk size of 3 latent frames, utilizing a global cache of 2 chunks and a local cache of 7 chunks. We refer readers to the supplementary materials for further details.

Table 1: Quantitative results for multi-shot video generation. The best results are highlighted in **boldface**, while the second best are underlined. Here, Sub., Bg., Cons., and Align. denote Subject, Background, Consistency, and Alignment, respectively.

Method	Architecture	FPS	Intra-shot Cons.		Inter-shot Cons.			Trans. Control ↑	Text Align. ↑	Aesthetic Quality ↑	Dynamic Degrees ↑
			Sub. ↑	Bg. ↑	Semantic ↑	Sub. ↑	Bg. ↑				
Mask2DiT [26]	Bidirectional	0.149	0.646	0.679	0.711	0.612	0.534	0.513	0.184	0.520	48.91
EchoShot [34]	Bidirectional	0.643	0.772	0.739	0.596	0.392	0.396	0.664	0.186	0.543	**65.92**
CineTrans [40]	Bidirectional	0.413	0.776	0.797	0.459	0.412	0.459	0.572	0.170	0.513	59.47
Self Forcing [12]	Causal	16.36	0.737	0.707	0.738	0.542	0.445	0.633	0.214	0.512	55.45
LongLive [42]	Causal	16.55	0.758	0.792	0.722	0.594	0.565	0.693	0.216	0.565	58.45
Rolling Forcing [20]	Causal	15.32	0.725	0.781	0.758	0.561	0.473	0.684	0.223	0.523	62.26
Infinity-RoPE [44]	Causal	16.37	0.752	0.738	0.622	0.453	0.407	0.715	0.209	0.513	63.40
ShotStream (Ours)	Causal	15.95	**0.825**	**0.819**	**0.762**	**0.654**	**0.645**	**0.978**	**0.234**	**0.571**	63.56

Evaluation Set. To comprehensively evaluate multi-shot video generation capabilities, following previous work [23, 34, 36, 40], we leverage Gemini 2.5 Pro [8] to generate 100 diverse multi-shot video prompts. To ensure a fair comparison, we tailor these text prompts to match the specific input style of each baseline model. These test prompts cover a wide range of themes, enabling a robust measurement of the models' ability to maintain consistency across different scenes.

Evaluation Metrics. Before computing metrics, we use the pretrained TransNet V2 [31] to detect shot boundaries in each video. We evaluate the model's multi-shot performance across five key dimensions: 1) *Intra-Shot Consistency*: Following HoloCine [23], we compute Subject Consistency using DINO [5] cosine similarities and Background Consistency using CLIP [27] similarities across frames. 2) *Inter-Shot Consistency*: Following MultiShotMaster [36], we isolate subjects and backgrounds from keyframes using YOLOv11 [16] and SAM [17], then assess consistency via DINOv2 [25]. Semantic Consistency is measured by computing the cosine similarity of ViCLIP [37] features extracted from each shot. 3) *Transition Control*: We utilize the Shot Cut Accuracy (SCA) metric [23] to evaluate the model's transition control capabilities by measuring the accuracy of cut counts and their temporal precision. 4) *Prompt Following*: We use Text Alignment [35, 36] to measure video-text consistency. 5) *Overall Quality*: We report Aesthetic Quality and Dynamic Degrees from VBench [13] to assess visual quality.

Baselines. We compare our model with relevant open-source video generation models of similar scale, including two categories: 1) *Bidirectional Multi-Shot Video Generation Model*: Mask2DiT [26] employs symmetric binary masks within the MMDiT architecture to isolate text annotations per segment while maintaining temporal coherence for multi-scene generation. EchoShot [34] targets portrait customization by using shot-aware position embeddings to model inter-shot relationships. CineTrans [40] designs a mask-based control mechanism for shot transitions. 2) *Autoregressive and Interactive Long Video Generation Model*: Baselines include Self Forcing [12], LongLive [42], Rolling Forcing [20],

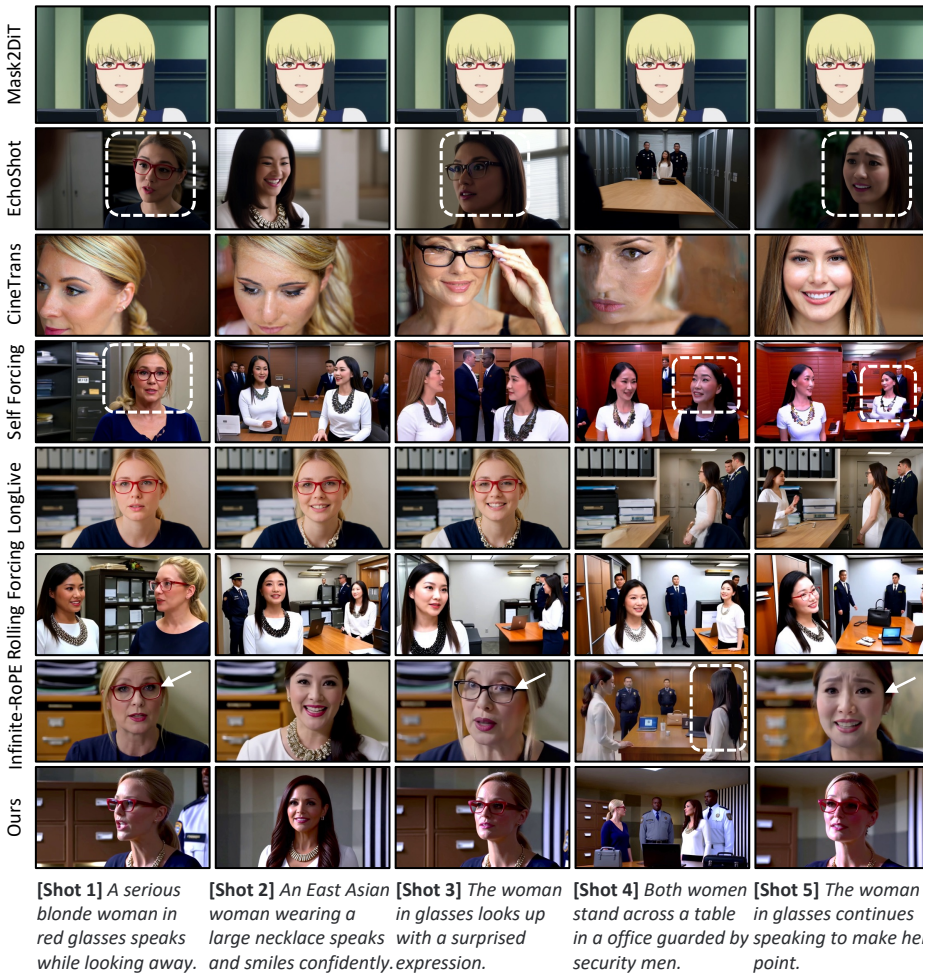

Fig. 5: Qualitative Comparison. We present the initial frames of each shot generated by all compared methods. Our approach not only adheres strictly to the prompts and maintains high visual coherence, but also produces natural transitions between shots.

and Infinity-RoPE [44]. LongLive [42] utilizes a KV-recache mechanism to refresh states with new prompts, enabling interactive generation. Infinity-RoPE [44] introduces a training-free RoPE Cut for multi-scene transitions in continuous rollouts. Although these autoregressive models can generate videos of several hundred frames, their multi-shot generation capabilities remain limited.

Quantitative Results. We validate ShotStream using the evaluation set described in Sec. 5.1. As shown in Table 1, our model outperforms competing methods across major metrics. It achieves the highest visual consistency while maintaining precise control over shot transitions, reflected by higher Consistency

Table 2: User Preference Rate. Participants select their preferred video from a randomized set of results generated by all methods. Multiple selections are allowed.

Method	Mask2DiT	EchoShot	CineTrans	Self Forcing	LongLive	Rolling Forcing	Infinity-RoPE	Ours
Visual Consistency	3.08%	12.31%	6.21%	1.54%	12.31%	15.38%	16.92%	**87.69**%
Prompt Following	0.83%	3.08%	1.54%	10.77%	16.15%	16.15%	14.62%	**76.15**%
Visual Quality	7.69%	18.46%	16.92%	10.77%	18.46%	23.08%	15.38%	**83.08**%

and Transition Control scores. Additionally, our method demonstrates superior prompt alignment for individual shots and higher overall aesthetic quality. We also evaluate inference efficiency of all methods using a single H200 GPU. Compared to bidirectional models, our approach yields more than 25× improvement in throughput (FPS). It also enables autoregressive long multi-shot video generation with minimal speed degradation relative to causal long-video models.

Qualitative Results. In Fig. 5, we provide a qualitative comparison on a complex, narrative-driven multi-shot prompt to illustrate the superiority of our method. Baseline methods, including Mask2DiT [26], CineTrans [40], Self Forcing [12], and Rolling Forcing [20] fail to generate shots that align with their respective prompts. While EchoShot [34] and Infinity-RoPE [44] successfully adapt to individual shot instructions, they struggle with inter-shot consistency. LongLive [42] confuses the identities of the two women appearing throughout the sequence. In contrast, our method faithfully adheres to multi-shot prompts while achieving high visual consistency and coherence with smooth transitions.

5.2 User Study

Due to the subjective nature of evaluating multi-shot video generation, we conducted a user study to compare different methods and validate the perceptual advantages of our proposed ShotStream. We randomly sampled 24 multi-shot prompts from the evaluation set (detailed in Sec. 5.1). During the test, participants are simultaneously presented with eight videos displayed in a randomized order: one generated by our method and seven from the competing baselines. Participants are asked to evaluate the videos from three aspects: Visual Consistency, Prompt Following, and Visual Quality. Multiple selections are allowed for each question. The user study involves 54 participants, and the results in Table 2 indicate that our method is consistently preferred by most users.

5.3 Ablation Studies

We perform ablation studies to validate the key design choices and training strategies of the bidirectional next-shot teacher and causal student models.

Bidirectional Next-Shot Teacher Model Design. As shown in Table 3, we validate our design choices across four key aspects: 1) *Context Frame Sampling Strategy:* To extract sparse context frames from historical shots, our model adopts a dynamic sampling strategy based on the number of historical shots to

Table 3: Ablation Study on the Bidirectional Next-Shot Teacher Model Design.

Method	Intra-shot Cons.		Inter-shot Cons.			Trans. Control ↑	Text Align. ↑	Aesthetic Quality ↑	Dynamic Degrees ↑
	Sub. ↑	Bg. ↑	Semantic ↑	Sub. ↑	Bg. ↑				
Context Frames Sampling Strategy									
First Only	0.789	0.793	0.671	0.618	0.612	0.956	0.212	0.502	61.55
First & Last	0.809	**0.827**	0.709	0.629	0.638	0.969	0.223	0.528	62.08
Dynamic (Ours)	**0.825**	0.819	**0.762**	**0.654**	**0.645**	**0.978**	**0.234**	**0.571**	**63.06**
Condition Frames Captioning Strategy									
Target Caption	0.804	0.818	0.681	0.609	0.572	0.937	0.194	0.422	62.86
Multi-Captions (Ours)	**0.825**	**0.819**	**0.762**	**0.654**	**0.645**	**0.978**	**0.234**	**0.571**	**63.06**
Condition Injection Mechanism									
Channel Concat	0.814	0.802	0.743	0.628	0.608	0.912	0.223	0.509	61.43
Frame Concat (Ours)	**0.825**	**0.819**	**0.762**	**0.654**	**0.645**	**0.978**	**0.234**	**0.571**	**63.06**
Training Strategy									
Full	0.816	0.810	0.749	0.631	0.624	0.969	0.227	0.546	60.85
Only 3D (Ours)	**0.825**	**0.819**	**0.762**	**0.654**	**0.645**	**0.978**	**0.234**	**0.571**	**63.06**

Table 4: Ablation Study on the Causal Student Model Design and Training Strategy.

Method	Intra-shot Cons.		Inter-shot Cons.			Trans. Control ↑	Text Align. ↑	Aesthetic Quality ↑	Dynamic Degrees ↑
	Sub. ↑	Bg. ↑	Semantic ↑	Sub. ↑	Bg. ↑				
Dual-Cache Distinction Strategy									
w/o Indicator	0.813	0.816	0.728	0.507	0.465	0.967	0.203	0.549	**63.09**
Learnable Emb.	0.811	0.809	0.737	0.518	0.588	0.972	0.204	0.523	61.19
RoPE Offset (Ours)	**0.825**	**0.819**	**0.762**	**0.654**	**0.645**	**0.978**	**0.234**	**0.571**	63.06
Causal Distillation Training Strategy									
Stage 1 Only	0.803	0.809	0.758	0.604	0.622	0.976	0.224	0.568	59.66
Stage 2 Only	0.819	0.814	0.704	0.583	0.547	0.975	0.218	0.467	52.86
Two Stage (Ours)	**0.825**	**0.819**	**0.762**	**0.654**	**0.645**	**0.978**	**0.234**	**0.571**	**63.06**

fulfill the conditional context budget. This aims to maximize the retention of historical information within a constrained budget. This dynamic approach outperforms naive baselines, such as sampling only the first frame or the first and last frames of each historical shot. 2) *Condition Frames Captioning Strategy*: To validate the necessity of injecting specific prompts for historical shots, we compare our approach against the widely used baseline [15, 22, 43] of using the same caption for both condition and target frames. Results indicate that injecting the corresponding prompts for condition frames is necessary and benefits in-context learning by effectively binding historical visual content with its respective text. 3) *Condition Injection Mechanism*: We evaluate our approach of injecting condition frames via temporal token concatenation against the commonly used channel concatenation method [7, 30]. This proves superior by allowing the 3D self-attention layers to directly model the relationships between condition and

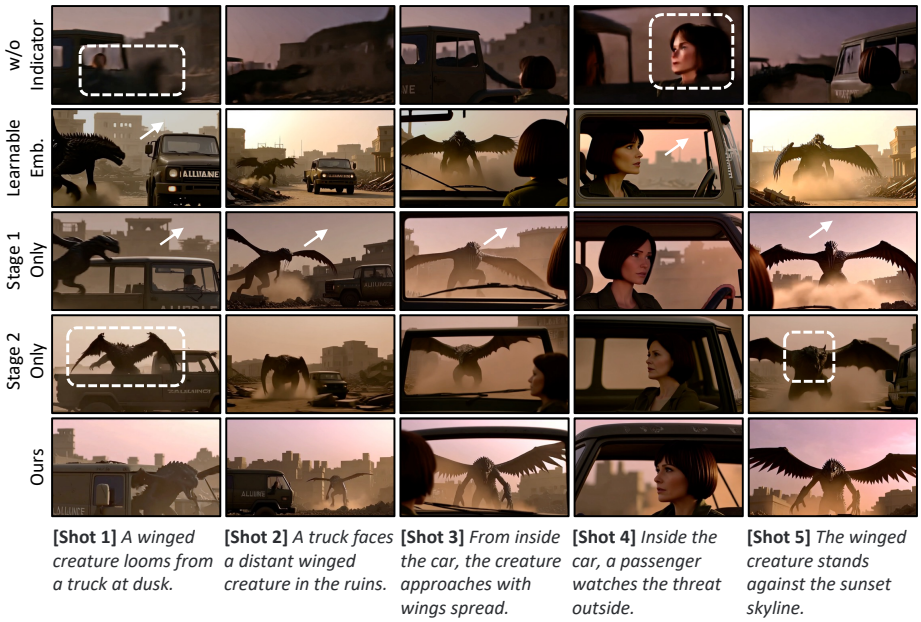

Fig. 6: Qualitative Ablation Results for the Causal Student Model. Please refer to the supplementary materials for video comparisons.

target tokens. 4) *Training Strategy*: We demonstrate that fine-tuning only the 3D self-attention layers outperforms full-parameter fine-tuning.

Causal Student Model Design. Table 4 presents ablations on the model design and distillation strategy of our causal model. 1) *Dual-Cache Distinction Strategy:* To justify the need to separate global from local caches, we compare our proposed RoPE offset (Row 3) against a baseline with no distinction (Row 1) and a variant using a learnable embedding applied to the target video's first chunk (Row 2). The results demonstrate that explicit distinction is essential (Row 1 vs. Row 3), and our training-free RoPE offset outperforms the learnable embedding approach (Row 2 vs. Row 3). 2) *Causal Distillation Training:* We evaluate our two-stage distillation strategy against single-stage baselines (Rows 4 and 5). Both stages prove indispensable: stage 1 establishes foundational next-shot generation capabilities, while stage 2 faithfully simulates inference to bridge the train-test gap. Furthermore, the qualitative results in Fig. 6 reinforce the necessity of both the RoPE offset and the two-stage distillation. Notably, the inter-shot self-forcing distillation significantly improves long-term consistency in visual style and color across the video (Stage 1 Only vs. Ours).

6 Conclusion

In this paper, we introduce ShotStream, a novel causal multi-shot video generation architecture that enables interactive long narrative storytelling while achieving 16 FPS on a single GPU. Our core contributions include reformulating the next-shot generation task for streaming, training a bidirectional next-shot teacher model, and distilling it into a causal architecture via a proposed two-stage distillation strategy. Additionally, we introduce a novel dual-cache memory mechanism to ensure visual consistency. Compared to existing bidirectional multi-shot models, ShotStream significantly reduces generation latency and supports streaming prompt inputs at runtime. This empowers users to interactively guide the narrative, adapting upcoming shots based on previously generated content. Furthermore, ShotStream advances the capabilities of autoregressive long video generation models by extending their ability to generate cohesive multi-shot sequences, paving the way for real-time, interactive, long-form storytelling.
Limitations and Future Work. While ShotStream is effective for autoregressive multi-shot video generation, we identify two primary limitations. First, we observe visual artifacts and inconsistencies when scenes and text prompts are highly complex. This primarily stems from the limited capacity of our backbone; because our current models are relatively small, we expect that scaling up the base model will improve performance and stability in challenging scenarios. Second, although our method is efficient, it still has room for acceleration to provide better interactive experiences. Techniques such as sparse attention [14, 52] and attention sink [30, 42] could be integrated into our model to achieve faster generation. We leave these extensions to future research.

Acknowledgements

The work is supported by the National Key R&D Program of China (No. 2025YFE0201300).

References

1. An, Z., Jia, M., Qiu, H., Zhou, Z., Huang, X., Liu, Z., Ren, W., Kahatapitiya, K., Liu, D., He, S., Zhang, C., Xiang, T., Yang, F., Belongie, S., Xie, T.: Onestory: Coherent multi-shot video generation with adaptive memory (2025), https://arxiv.org/abs/2512.07802
2. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), https://openai.com/research/video-generation-models-as-world-simulators
3. Bruce, J., Dennis, M.D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al.: Genie: Generative interactive environments. In: Forty-first International Conference on Machine Learning (2024)
4. Cai, S., Yang, C., Zhang, L., Guo, Y., Xiao, J., Yang, Z., Xu, Y., Yang, Z., Yuille, A., Guibas, L., Agrawala, M., Jiang, L., Wetzstein, G.: Mixture of contexts for long video generation (2025), https://arxiv.org/abs/2508.21058

5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
6. Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., Xiong, W., Wang, W., Pang, N., Kang, K., Xu, Z., Jin, Y., Liang, Y., Song, Y., Zhao, P., Xu, B., Qiu, D., Li, D., Fei, Z., Li, Y., Zhou, Y.: Skyreels-v2: Infinite-length film generative model (2025), https://arxiv.org/abs/2504.13074
7. Chen, Y., Guo, S., Jin, R., Yang, T., Cai, X., Luo, Y., Yang, M., Yu, M., Xu, L., Xue, T.: Anyrecon: Arbitrary-view 3d reconstruction with video diffusion model. arXiv preprint arXiv:2604.19747 (2026)
8. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
9. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Self-forcing++: Towards minute-scale high-quality video generation. arXiv preprint arXiv:2510.02283 (2025)
10. Guo, Y., Yang, C., Yang, Z., Ma, Z., Lin, Z., Yang, Z., Lin, D., Jiang, L.: Long context tuning for video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17281–17291 (October 2025)
11. He, J., Liu, H., Li, J., Huang, Z., Yu, Q., Ouyang, W., Liu, Z.: Cut2next: Generating next shot via in-context tuning. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–11 (2025)
12. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
13. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
14. Jia, W., Lu, Y., Huang, M., Wang, H., Huang, B., Chen, N., Liu, M., Jiang, J., Mao, Z.: Moga: Mixture-of-groups attention for end-to-end long video generation (2025), https://arxiv.org/abs/2510.18692
15. Ju, X., Ye, W., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Xu, Q.: Fulldit: Multi-task video generative foundation model with full attention. arXiv preprint arXiv:2503.19907 (2025)
16. Khanam, R., Hussain, M.: Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725 (2024)
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
18. Kuaishou: Kling video model. https://kling.kuaishou.com/en (2024)
19. Liu, J., Li, S., Fang, Z., Li, X., Zhou, Y., Meng, Z., Zhang, Z., Luo, Y., Zhang, G., Liu, Y.S., et al.: Omnidirector: General multi-shot camera cloning without cross-paired data. arXiv preprint arXiv:2606.13432 (2026)
20. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time. arXiv preprint arXiv:2509.25161 (2025)
21. Lu, Y., Zeng, Y., Li, H., Ouyang, H., Wang, Q., Cheng, K.L., Zhu, J., Cao, H., Zhang, Z., Zhu, X., et al.: Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. arXiv preprint arXiv:2512.04678 (2025)

22. Luo, Y., Shi, X., Bai, J., Xia, M., Xue, T., Wang, X., Wan, P., Zhang, D., Gai, K.: Camclonemaster: Enabling reference-based camera control for video generation. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–10 (2025)
23. Meng, Y., Ouyang, H., Yu, Y., Wang, Q., Wang, W., Cheng, K.L., Wang, H., Li, Y., Chen, C., Zeng, Y., Shen, Y., Qu, H.: Holocine: Holistic generation of cinematic multi-shot long video narratives (2025), https://arxiv.org/abs/2510.20822
24. Ning, M., Li, M., Su, J., Salah, A.A., Ertugrul, I.O.: Elucidating the exposure bias in diffusion models. arXiv preprint arXiv:2308.15321 (2023)
25. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
26. Qi, T., Yuan, J., Feng, W., Fang, S., Liu, J., Zhou, S., He, Q., Xie, H., Zhang, Y.: Mask^ 2dit: Dual mask-based diffusion transformer for multi-scene long video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18837–18846 (2025)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
28. RunwayML: Gen-3 alpha. https://runwayml.com/research/introducing-gen-3-alpha (2024)
29. Schmidt, F.: Generalization in generation: A closer look at exposure bias. arXiv preprint arXiv:1910.00292 (2019)
30. Shin, J., Li, Z., Zhang, R., Zhu, J.Y., Park, J., Shechtman, E., Huang, X.: Motionstream: Real-time video generation with interactive motion controls. arXiv preprint arXiv:2511.01266 (2025)
31. Soucek, T., Lokoc, J.: Transnet v2: An effective deep network architecture for fast shot transition detection. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 11218–11221 (2024)
32. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al.: Magi-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211 (2025)
33. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
34. Wang, J., Sheng, H., Cai, S., Zhang, W., Yan, C., Feng, Y., Deng, B., Ye, J.: Echoshot: Multi-shot portrait video generation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
35. Wang, Q., Luo, Y., Shi, X., Jia, X., Lu, H., Xue, T., Wang, X., Wan, P., Zhang, D., Gai, K.: Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–10 (2025)
36. Wang, Q., Shi, X., Li, B., Bian, W., Liu, Q., Lu, H., Wang, X., Wan, P., Gai, K., Jia, X.: Multishotmaster: A controllable multi-shot video generation framework. arXiv preprint arXiv:2512.03041 (2025)
37. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv preprint arXiv:2307.06942 (2023)

38. Wang, Y., Xiong, T., Zhou, D., Lin, Z., Zhao, Y., Kang, B., Feng, J., Liu, X.: Loong: Generating minute-level long videos with autoregressive language models. arXiv preprint arXiv:2410.02757 (2024)
39. Weissenborn, D., Täckström, O., Uszkoreit, J.: Scaling autoregressive video models. arXiv preprint arXiv:1906.02634 (2019)
40. Wu, X., Gao, B., Qiao, Y., Wang, Y., Chen, X.: Cinetrans: Learning to generate videos with cinematic transitions via masked diffusion models. arXiv preprint arXiv:2508.11484 (2025)
41. Xiao, J., Yang, C., Zhang, L., Cai, S., Zhao, Y., Guo, Y., Wetzstein, G., Agrawala, M., Yuille, A., Jiang, L.: Captain cinema: Towards short movie generation (2025), https://arxiv.org/abs/2507.18634
42. Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., Li, M., Xie, E., Chen, Y., Lu, Y., et al.: Longlive: Real-time interactive long video generation. arXiv preprint arXiv:2509.22622 (2025)
43. Ye, Z., He, X., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Chen, Q., Luo, W.: Unic: Unified in-context video editing. arXiv preprint arXiv:2506.04216 (2025)
44. Yesiltepe, H., Meral, T.H.S., Akan, A.K., Oktay, K., Yanardag, P.: Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. arXiv preprint arXiv:2511.20649 (2025)
45. Yi, J., Jang, W., Cho, P.H., Nam, J., Yoon, H., Kim, S.: Deep forcing: Training-free long video generation with deep sink and participative compression. arXiv preprint arXiv:2512.05081 (2025)
46. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, B.: Improved distribution matching distillation for fast image synthesis. Advances in neural information processing systems **37**, 47455–47487 (2024)
47. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024)
48. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast causal video generators. arXiv e-prints pp. arXiv–2412 (2024)
49. Zhang, K., Jiang, L., Wang, A., Fang, J.Z., Zhi, T., Yan, Q., Kang, H., Lu, X., Pan, X.: Storymem: Multi-shot long video storytelling with memory (2025), https://arxiv.org/abs/2512.19539
50. Zheng, M., Xu, Y., Huang, H., Ma, X., Liu, Y., Shu, W., Pang, Y., Tang, F., Chen, Q., Yang, H., et al.: Videogen-of-thought: Step-by-step generating multi-shot video with minimal manual intervention. arXiv preprint arXiv:2412.02259 (2024)
51. Zhou, Y., Zhou, D., Cheng, M.M., Feng, J., Hou, Q.: Storydiffusion: Consistent self-attention for long-range image and video generation. Advances in Neural Information Processing Systems **37**, 110315–110340 (2024)
52. Zhuang, J., Guo, S., Cai, X., Li, X., Liu, Y., Yuan, C., Xue, T.: Flashvsr: Towards real-time diffusion-based streaming video super-resolution. arXiv preprint arXiv:2510.12747 (2025)