CausalCine: Real-Time Autoregressive Generation for Multi-Shot Video Narratives

Yihao Meng[1,2,*] **Zichen Liu**[1,2,*] **Hao Ouyang**[2] **Qiuyu Wang**[2]
Ka Leong Cheng[2] **Yue Yu**[1,2] **Hanlin Wang**[1,2] **Xing Zhu**[2]
Yujun Shen[2] **Qifeng Chen**[1] **Huamin Qu**[1]
[1]The Hong Kong University of Science and Technology
[2]Ant Group
ymengas@connect.ust.hk [*]Equal contribution

Abstract

Autoregressive video generation aims at real-time, open-ended synthesis. Yet, cinematic storytelling is not merely the endless extension of a single scene; it requires progressing through evolving events, viewpoint shifts, and discrete shot boundaries. Existing autoregressive models often struggle in this setting. Trained primarily for short-horizon continuation, they treat long sequences as extended single shots, inevitably suffering from motion stagnation and semantic drift during long rollouts. To bridge this gap, we introduce **CausalCine**, an interactive autoregressive framework that transforms multi-shot video generation into an online directing process. CausalCine generates causally across shot changes, accepts dynamic prompts on the fly, and reuses context without regenerating previous shots. To achieve this, we first train a causal base model on native multi-shot sequences to learn complex shot transitions prior to acceleration. We then propose Content-Aware Memory Routing (CAMR), which dynamically retrieves historical KV entries according to attention-based relevance scores rather than temporal proximity, preserving cross-shot coherence under bounded active memory. Finally, we distill the causal base model into a few-step generator for real-time interactive generation. Extensive experiments demonstrate that CausalCine significantly outperforms autoregressive baselines and approaches the capability of bidirectional models while unlocking the streaming interactivity of causal generation. Project page and video demonstrations are available at https://yihao-meng.github.io/CausalCine/.

1 Introduction

Recent diffusion video models achieve impressive visual fidelity [37, 41, 13], but their bidirectional attention makes long, interactive generation expensive. Autoregressive generation with KV caching offers a natural alternative for streaming video synthesis [18, 2], yet existing causal video models are still largely trained and evaluated as short-horizon continuation systems [55, 17, 62, 51, 26]. When rolled out beyond a single local motion pattern, they often stagnate, loop, or drift semantically [4]. Cinematic long-form video, however, is not merely an extended single shot. It requires evolving events, viewpoint changes, discrete shot boundaries, and persistent story context.

In this work, we study interactive autoregressive multi-shot video generation, where a model generates videos causally across shot changes, accepts new prompts during generation, and reuses relevant long-range context without regenerating previous shots. This setting exposes a key limitation of short-clip autoregressive training: the model must not only continue local motion, but also introduce new content at requested shot changes, follow newly appended prompts, and determine which information from earlier shots should remain accessible.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

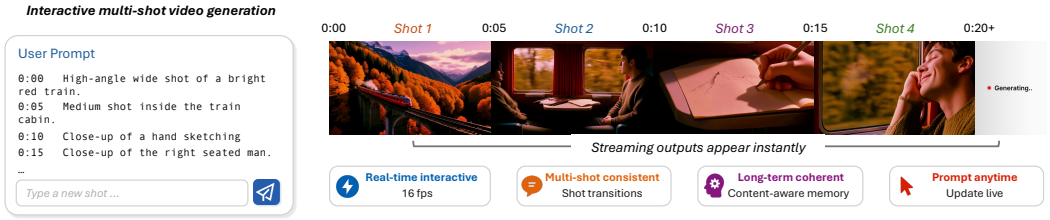


Figure 1: Real-time interactive multi-shot generation. CausalCine streams video chunks causally, accepts new shot-level prompts during rollout, and reuses content-aware KV memory so later shots can recall earlier content without regenerating previous video.

Our first observation is that long-form causal behavior should be learned before acceleration. Instead of directly distilling a bidirectional diffusion model into a fast autoregressive generator [55, 17], we first train a full-step causal multi-shot base model on native long-form sequences with teacher forcing. The model observes shot boundaries, changing prompts, and long-range entity reappearance under the same causal dependency structure used at inference with KV caching. We find that high-quality native multi-shot data substantially reduces the usual teacher-forcing rollout gap, yielding a causal base model that can perform stable long rollouts and synthesize new content across shot transitions.

Autoregressive multi-shot generation also poses a greater challenge for KV memory. In single-scene continuation, fixed anchors or sliding windows can preserve local appearance and motion continuity [51, 18, 26]. However, when generation must introduce new content, viewpoints, or environments, useful context is no longer determined by temporal proximity or fixed frame positions. The model may need to recall a character from the distant past, ignore the immediately preceding scene, or combine semantic cues from multiple earlier shots. We therefore introduce Content-Aware Memory Routing (CAMR), which selects historical KV entries by content relevance rather than fixed temporal position. CAMR retrieves useful long-range context and maintains a streamlined memory representation, improving cross-shot coherence without sacrificing causal generation.

Finally, we distill the causal multi-shot base model into a few-step generator for real-time interactive synthesis. Because causality and multi-shot structure have already been learned by the full-step model, Distribution Matching Distillation (DMD) [54, 53] can focus on trajectory compression while preserving visual quality and cross-shot consistency. The resulting model generates videos chunk by chunk with KV caching, supports prompt updates during generation, and continues a sequence without recomputing previous shots.

The resulting system enables real-time online directing for long-form video generation. Rather than rendering a complete video offline, CausalCine streams video causally: users can start from an initial shot, issue new prompts during rollout, introduce new events or viewpoints, and continue generation without recomputing previous shots. Importantly, this capability is demonstrated at practical model scale. We build CausalCine on a 14B-parameter video generator and run it with streaming KV caching on 8 NVIDIA H200 GPUs at 16 FPS. This makes interactive multi-shot generation possible in real time, while preserving long-range semantic memory across shot boundaries. Experiments show that CausalCine substantially outperforms autoregressive baselines in shot-level quality, prompt alignment, identity preservation, and transition structure, and approaches the visual quality of bidirectional models while retaining the efficiency and interactivity unique to causal generation.

2 Related Works

2.1 Autoregressive Video Generation

Autoregressive video generation factorizes a video into sequentially generated frames or chunks, making it naturally suited for long-horizon rollout, KV-cache reuse, and interactive continuation [2, 18, 23, 38, 27]. Recent autoregressive video models often start from pretrained diffusion models, then make them causal so that videos can be generated chunk by chunk [55, 17, 62, 51, 26, 7]. CausVid [55] distills bidirectional diffusion into a few-step causal model for low-latency streaming, while Self Forcing [17] and Causal Forcing [62] reduce train–test mismatch by supervising the model on its own rollout distribution with distribution matching distillation [54, 53]. Long-context AR systems further extend generation through rolling caches, local windows, fixed anchors, or runtime

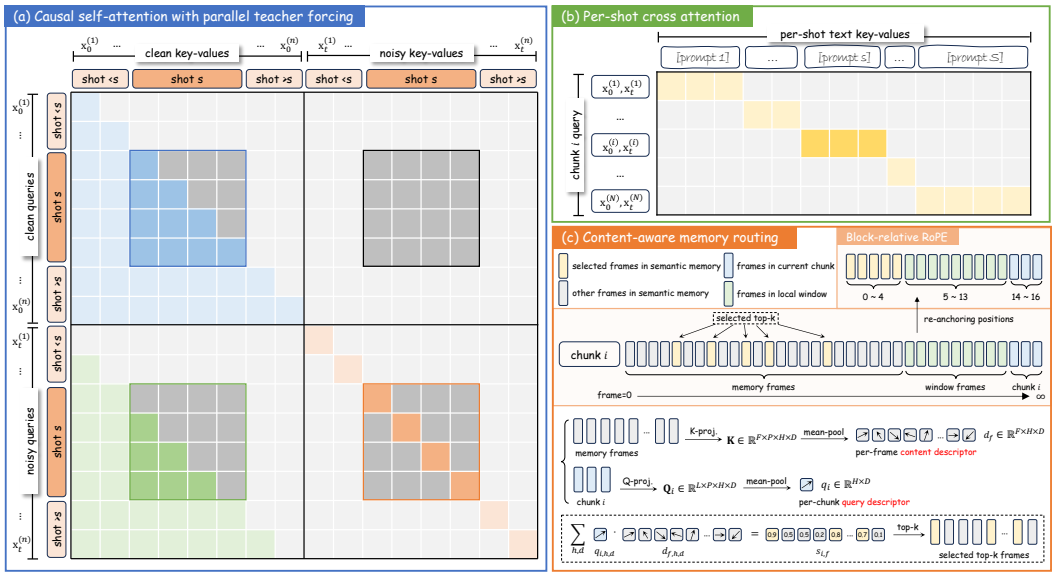


Figure 2: Overview of CausalCine. (a) A 2N-segment teacher-forcing layout trains causal multi-shot dependencies in one parallel forward pass. (b) Per-shot cross-attention routes each chunk to its active shot prompt. (c) Content-Aware Memory Routing retrieves relevant historical KV entries and applies Block-Relative RoPE to keep positional phases within the training range during long rollouts.

prompt updates [51, 26, 52, 8]. However, these methods are primarily designed for single-scene continuation, where long video generation is treated as extending a local motion pattern. We study autoregressive generation in the multi-shot setting, where the model causally introduces new shots, prompts, and events while preserving long-range story context.

2.2 Multi-Shot Video Generation

Multi-shot video generation aims to synthesize coherent long videos with multiple shots, scene transitions, and evolving story structure. Existing approaches often decompose the task into scripts, shots, or keyframes, and generate each segment with a short-video model [1, 16, 28, 44, 50, 58, 48, 60, 61]. These methods provide explicit control over story planning, but cross-shot consistency must be recovered through separate linking or refinement stages. More recent holistic methods model multiple shots jointly inside a unified diffusion process [31, 3, 22, 34, 42, 12], improving global consistency by allowing all shots to interact during generation. However, their bidirectional formulation requires joint generation over all shots, leading to quadratic cost with video length and limiting online interaction. In contrast, CausalCine generates multi-shot videos autoregressively, allowing new prompts to be appended on the fly without recomputing previous content.

2.3 Memory in Video Generation Models

Memory mechanisms are widely used to extend video generation beyond the local temporal window. Streaming AR models typically retain recent frames together with fixed anchors or sink tokens from the sequence beginning [47, 51], while other methods compress history into compact representations or maintain multi-scale short- and long-term memory [57, 10, 14, 15]. More recent work explores adaptive memory, retrieving history based on camera pose, field-of-view overlap, 3D scene structure, or content relevance [49, 56, 24, 3, 21, 11]. Inspired by these directions, we integrate content-aware memory retrieval directly into the visual KV cache, and show that such adaptive memory is effective for the more challenging setting of few-step causal multi-shot generation.

3 Method

We organize our framework around the design rationale that *causality and multi-shot structure should be learned before step compression*. Starting from a pretrained bidirectional flow-matching video

diffusion model, we (i) tune it into a full-step causal multi-shot generator with parallel teacher forcing on long cinematic videos (Sec. 3.2); (ii) replace its temporal positional heuristics with a content-aware memory router shared by training and inference (Sec. 3.3); and (iii) distill the resulting full-step causal model into a four-step generator for interactive synthesis (Sec. 3.4).

3.1 Preliminaries

Flow-matching video diffusion. We operate in the video VAE latent space, where a clean video latent $\mathbf{x}_0 \in \mathbb{R}^{F\times C\times H\times W}$ and Gaussian noise $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$ are interpolated as $\mathbf{x}_t = (1-\sigma_t)\mathbf{x}_0 + \sigma_t\boldsymbol{\epsilon}$ under a shifted schedule [9]. A DiT [33] velocity field $v_\theta(\mathbf{x}_t, t, \mathbf{c})$ is trained with the rectified flow-matching loss

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\big\|v_\theta(\mathbf{x}_t,t,\mathbf{c}) - (\boldsymbol{\epsilon}-\mathbf{x}_0)\big\|^2, \tag{1}$$

and sampling integrates $\mathrm{d}\mathbf{x}/\mathrm{d}t = v_\theta$ with a few-step Euler solver.

Distribution matching distillation. DMD [54, 53] compresses a pretrained teacher into a few-step student G_ϕ by minimizing a reverse KL between the student and teacher distributions at every noise level t, yielding the implicit gradient

$$\nabla_\phi\mathcal{L}_{\text{DMD}} = \mathbb{E}_t\big[(s_{\text{fake}}(\mathbf{x}_t,t) - s_{\text{real}}(\mathbf{x}_t,t))\,\partial_\phi G_\phi\big], \tag{2}$$

where s_{real} is predicted by the frozen teacher and s_{fake} is predicted by an auxiliary score network co-trained with flow matching on the student's rollouts. We use this formulation in Sec. 3.4 and augment it with adversarial regularization.

3.2 Long Multi-Shot Causal Tuning

This stage converts a pretrained bidirectional video diffusion transformer into a causal generator whose training already covers the distribution of cinematic shot transitions, through a unified teacher-forcing regime on long multi-shot videos.

Causal chunk-wise formulation. We factorize a long video latent along the temporal axis into N contiguous *chunks* $\{\mathbf{x}^{(1)},\dots,\mathbf{x}^{(N)}\}$ with $\mathbf{x}^{(i)}\in\mathbb{R}^{L\times C\times H\times W}$. A chunk is the *unit of autoregression*, not a narrative unit; in our experiments $L{=}3$ latent frames ($\approx$12 video frames), and frame-wise AR is the special case $L{=}1$. The joint distribution factorizes causally,

$$p_\theta(\mathbf{x}^{(1:N)} \mid \mathbf{c}_{1:N}) = \prod_{i=1}^{N} p_\theta(\mathbf{x}^{(i)} \mid \mathbf{x}^{(<i)}, \mathbf{c}_i). \tag{3}$$

A multi-shot video consists of S contiguous shots with prompts $\{\mathbf{c}_{(s)}\}_{s=1}^{S}$ separated by latent-frame boundaries $\mathcal{B} = \{b_1,\dots,b_{S-1}\}$. The text condition for chunk i is therefore *shot-indexed*, $\mathbf{c}_i = \mathbf{c}_{(\pi(i))}$, where $\pi(i)\in\{1,\dots,S\}$ is the shot containing chunk i. At a shot boundary the prompt $\mathbf{c}_i$ changes; the generated chunk $\mathbf{x}^{(i)}$ is then expected to faithfully reflect the new prompt rather than extrapolate the previous shot, a regime in which short clip-trained AR models tend to collapse onto static or looping content [4].

Parallel teacher forcing with $2N$-segment packing. A step-by-step rollout of Eq. (3) during training is prohibitive in time and memory. Following teacher-forcing training [17, 62], we pack, for each video, a single $2N$-segment input of clean and noisy copies of all chunks:

$$\mathbf{X}_{\text{TF}} = \big[\underbrace{\mathbf{x}_0^{(1)},\dots,\mathbf{x}_0^{(N)}}_{\text{clean context}},\ \underbrace{\mathbf{x}_t^{(1)},\dots,\mathbf{x}_t^{(N)}}_{\text{noisy queries}}\big]. \tag{4}$$

Clean segments carry timestep 0; all noisy segments share a single sampled $t\sim p(\sigma)$, keeping the loss aligned across chunks. The block-sparse self-attention mask (Fig. 2(a)) has four quadrants: (a) clean$\to$clean is causal, where each clean chunk attends to itself and all preceding clean chunks; (b) noisy$\to$clean allows each noisy chunk to attend only to preceding clean chunks; (c) noisy$\to$noisy is restricted to the diagonal, ruling out leakage from future noisy chunks; and (d) clean$\to$noisy is fully masked. The flow-matching loss is computed on the noisy half:

$$\mathcal{L}_{\text{tune}} = \mathbb{E}_{t,\mathbf{X}_{\text{TF}}}\frac{1}{N}\sum_{i=1}^{N}\big\|v_\theta(\mathbf{X}_{\text{TF}};t,\mathcal{M})_{[N+i]} - (\boldsymbol{\epsilon}^{(i)}-\mathbf{x}_0^{(i)})\big\|^2. \tag{5}$$

This layout exposes the causal visibility pattern that the model uses with a KV cache at inference, while replacing sequential rollout with a single parallel forward pass. In practice, the noisy→clean quadrant is further sparsified into a local window plus content-routed long memory, as in Sec. 3.3.

Per-shot text conditioning. As shown in Fig. 2(a), given shot boundaries $\mathcal{B}$, both segments of chunk i in the packed layout (the clean context $\mathbf{x}_0^{(i)}$ and the noisy query $\mathbf{x}_t^{(i)}$) are conditioned on the same shot prompt $\mathbf{c}_{(\pi(i))}$ via segment-level cross-attention; cross-attention between segments is forbidden, so each chunk only sees its own shot's prompt tokens. This explicit shot-indexed routing ties shot-boundary prompt changes to visual transitions.

Scaling to long cinematic videos. Short clips rarely span shot boundaries, failing to supervise transition dynamics or long range entity correlation, the very essence of cinematic videos. To learn these behaviors, we train natively on long multi-shot sequences of $\sim$15 s ($\approx$241 video frames). This long-form supervision provides the critical signals needed for the causal model to actively introduce new scenes and preserve identities across cuts, rather than merely extrapolating a single shot. To make this extensive context tractable, our $2N$-packing trains all targets in a single parallel pass, while FSDP [59] and sequence-parallel attention [20] absorb the $\mathcal{O}(NL)$ memory footprint.

3.3 Content-Aware Memory Routing

Long rollouts require compressing the growing KV cache into a bounded attention buffer. Prior AR video generators typically use position-defined memory, such as a local window plus first-frame sink tokens [47, 62], which is fragile when multi-shot generation introduces new or reappearing content far from the opening frame. We instead augment the local window with content-addressable memory: each attention layer retrieves history frames whose keys best match the current query. The same routing module is used in TF training and AR inference, as shown in Fig. 2 (c).

Frame-level, chunk-shared routing. Let $\mathbf{K} \in \mathbb{R}^{F\times P\times H\times D}$ stack the cached keys of F history latent frames, where P is the number of spatial tokens per frame, H the number of heads, and D the head dimension. Following prior token-level sparse routing in language and video models [29, 3], where mean-pooled keys have been shown to provide effective retrieval signals, for every cached frame f we store a compact *content descriptor* obtained by mean-pooling its key over spatial tokens,

$$\mathbf{d}_f \;=\; \frac{1}{P}\sum_{p=1}^{P} \mathbf{K}_{f,p,:,:} \;\in\; \mathbb{R}^{H\times D}. \tag{6}$$

For the current chunk $\mathbf{x}^{(i)}$ we form a query descriptor $\mathbf{q}_i \in \mathbb{R}^{H\times D}$ in the same way, mean-pooling the chunk's queries over both its L frames and P spatial tokens, so that all L frames in chunk i share one routing decision. We score every out-of-window history frame by a head-aggregated dot product,

$$s_{i,f} \;=\; \sum_{h,d} \mathbf{q}_{i,h,d}\, \mathbf{d}_{f,h,d}, \tag{7}$$

and select the top-k frames. Letting $\mathcal{W}_i$ denote the local window of W chunks preceding i and $\mathcal{H}_i$ the out-of-window history, the effective receptive field of chunk i is

$$\mathcal{R}_i \;=\; \underbrace{\text{Top-}k\big(\{s_{i,f}\}_{f\in\mathcal{H}_i}\big)}_{\text{semantic memory}} \cup \underbrace{\mathcal{W}_i}_{\text{local window}} \cup \{\text{current chunk}\}. \tag{8}$$

We use $W{=}3$ chunks and $k{=}5$ frames throughout. The routing is model-adaptive but parameter-free. Although the top-k selection is not differentiable, scores are computed from the learned query/key representations. Routing is applied to self-attention only; cross-attention to text remains as in Sec. 3.2.

Block-Relative RoPE. Content-based routing may retrieve frames beyond the training horizon F_{train}, e.g.,1000th frame in a minute-long rollout. Applying 3D RoPE at these global positions exposes attention to unseen phases and may cause severe visual artifacts.

We avoid this by re-anchoring positions after retrieval. Keys are stored unrotated in the cache; after top-k selection, RoPE is applied to the selected memory, local window, and current chunk using

compact block-relative positions:

$$\underbrace{[0,\ldots,k{-}1]}_{\text{memory}} \,\|\, \underbrace{[k,\ldots,k{+}WL{-}1]}_{\text{window}} \,\|\, \underbrace{[k{+}WL,\ldots,k{+}(W{+}1)L{-}1]}_{\text{current}}, \tag{9}$$

whose span is $k + (W{+}1)L \le F_{\text{train}}$ by construction ($5 + 4{\cdot}3 = 17 \ll 61$ in our setting). Since the same cached key may receive different relative positions for different queries, keys cannot be rotated once at write time. This Block-Relative RoPE keeps all attention phases within the training-range envelope regardless of rollout length.

3.4 Few-Step Causal Distillation

With the causal multi-shot model from Sec. 3.2 and the memory router from Sec. 3.3, we distill the many-step flow-matching teacher into a four-step autoregressive generator G_ϕ using Distribution Matching Distillation (DMD) [54, 53] and an adversarial objective. The distilled student preserves the causal chunk-wise architecture, per-shot conditioning, without modification.

Teacher Forcing causal ODE initialization. Before self-forced DMD, we initialize the student via causal ODE distillation [62]. Given a ground-truth history $\mathbf{x}_{\text{gt}}^{(<i)}$ and shot prompt $\mathbf{c}_i$, we generate a teacher PF-ODE trajectory $\mathbf{z}_\tau^{(i)}$ from noise $\boldsymbol{\epsilon}^{(i)}$ (subsampled to 4 steps $\tau \in \mathcal{S}$ from a 48-step solver). We train the student to predict the teacher's final denoised output $\mathbf{z}_0^{(i)}$ by minimizing:

$$\mathcal{L}_{\text{init}} = \mathbb{E}_{i,\tau\sim\mathcal{S}} \left\| \hat{\mathbf{x}}_{0,\phi}\left(\mathbf{z}_\tau^{(i)}, \mathbf{x}_{\text{gt}}^{(<i)}, \tau, \mathbf{c}_i\right) - \mathbf{z}_0^{(i)} \right\|_2^2. \tag{10}$$

This aligns the few-step student with the teacher's causal visibility pattern, which is crucial for preventing unstable targets during the subsequent self-forced training where the teacher's scores are queried on the student's own long-horizon rollouts.

Distribution matching distillation with adversarial regularization. We further refine G_ϕ under a self-forcing framework [17]: each update starts from the student's own causal rollout $\tilde{\mathbf{x}}_{0,\phi}$ using the inference KV cache and memory routing. After perturbing it to $\tilde{\mathbf{x}}_{t,\phi}$, we apply the DMD gradient (eq. (2)). The frozen real denoiser T_ψ and the flow-matching-updated auxiliary fake denoiser T_{ϕ^-} are initialized from our tuned multi-shot model.

To reduce sequence-level drift in long rollouts, we follow APT [25] and attach a lightweight GAN head D_η to the intermediate features of T_{ϕ^-}. Let $d_\eta(\mathbf{y}_t) = D_\eta(F_{\phi^-}(\mathbf{y}_t, t, \mathbf{c}))$ denote the logit. We optimize the standard logistic adversarial loss:

$$\mathcal{L}_D = \mathbb{E}_{\mathbf{x}_0}[f(-d_\eta(\mathbf{x}_t))] + \mathbb{E}_{\tilde{\mathbf{x}}_{0,\phi}}[f(d_\eta(\tilde{\mathbf{x}}_{t,\phi}))], \quad \mathcal{L}_G = \mathcal{L}_{\text{DMD}} + \lambda_{\text{adv}}\mathbb{E}_{\tilde{\mathbf{x}}_{0,\phi}}[f(-d_\eta(\tilde{\mathbf{x}}_{t,\phi}))], \tag{11}$$

where $f(u) = \log(1 + \exp(u))$. The generator is trained with $\mathcal{L}_G$ and the discriminator with $\mathcal{L}_D$, effectively penalizing drift in camera motion and subject framing.

4 Experiments

Implementation Details. We build our autoregressive framework on Wan2.1-T2V-14B [41] to generate videos at resolution 832×480. The causal base model is trained with chunk-wise teacher forcing on 100k long multi-shot videos, where each chunk contains three latent frames. The training process is conducted on 64 NVIDIA H800 GPUs. At inference time, the model generates chunks sequentially with KV caching. Our distilled student uses four denoising steps and inherits the same per-shot text routing and memory mechanism as the causal base.

Evaluation Protocol. Following Meng et al. [31], we use Gemini 2.5 Pro [40, 6] to build a 100-prompt multi-shot benchmark. Each prompt contains a global story description, five shot-level captions, and target shot-cut locations, covering character reappearance, scene changes, shot-reverse-shot interactions, viewpoint changes, and long temporal gaps. Following VBench [19], we evaluate visual quality, prompt following, temporal consistency, long-range consistency, and shot structure. Specifically, we report LAION aesthetic score [36], shot-level ViCLIP text-video similarity [46, 45], within-shot subject/background consistency using DINO [5] and CLIP [35], inter-shot character consistency using DINOv2 [32] on matched pairs, and shot-cut accuracy (SCA) [31] by matching TransNetV2 [39]-detected cuts to target boundaries. To ensure a fair comparison, all baselines generate videos under identical settings as ours, using the same set of prompts, resolution, and length.

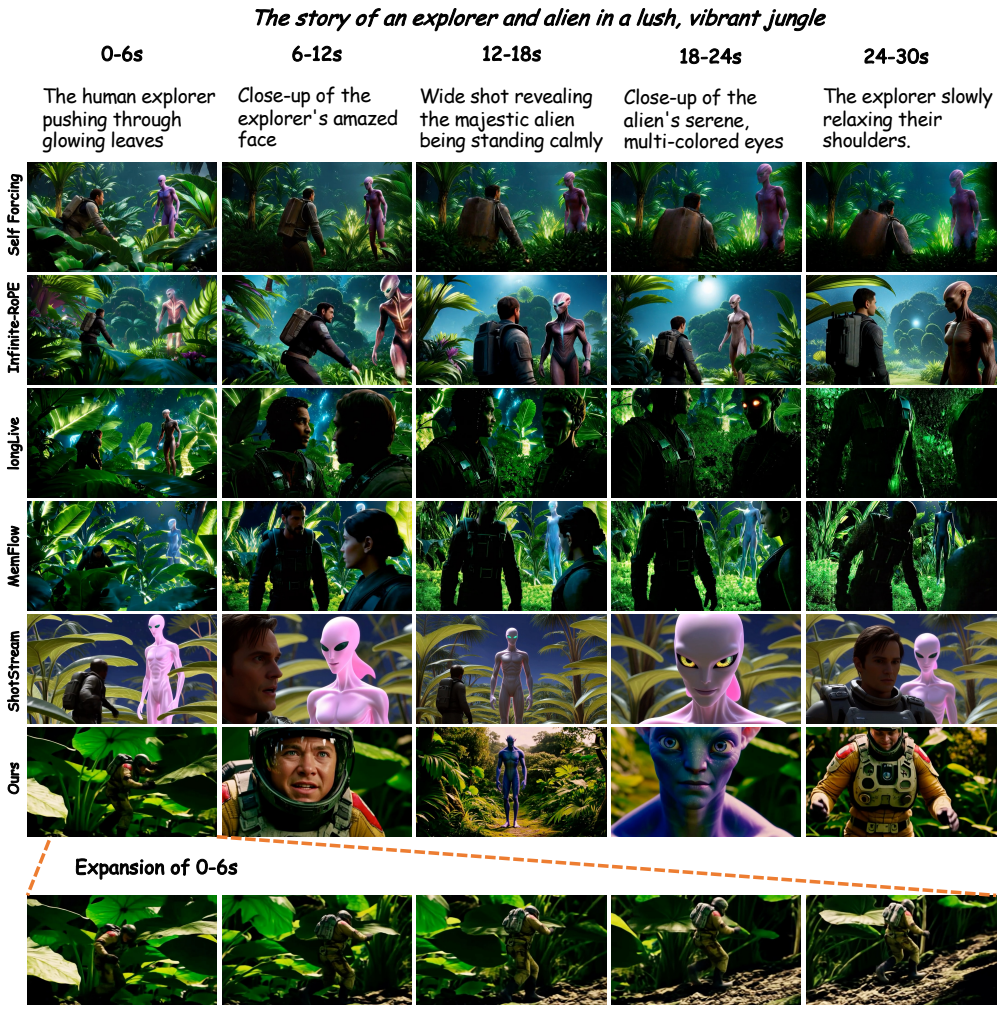


Figure 3: Comparison with autoregressive and streaming long-video baselines. Existing autoregressive methods often remain tied to the initial scene, repeat similar compositions, or miss requested viewpoint changes. CausalCine better follows the shot progression while preserving subjects across shots; the expanded first shot shows coherent intra-shot motion.



Figure 4: Our causal generator produces results comparable to bidirectional baselines.

Table 1: Comparison with autoregressive video generation baselines. Best values per column are in **bold** and second best are underlined. Our method achieves the best overall performance.

Method	Aesthetic ↑	Text Align. ↑	Subject Consistency ↑	Background Consistency ↑	SCA. ↑
Self-Forcing	0.6228	0.1395	0.9668	**0.9717**	0.5052
Infinity-RoPE	0.6225	0.1716	0.8609	0.9091	0.7842
LongLive	0.6198	0.1552	0.9319	0.9487	0.5021
MemFlow	0.6139	0.1587	0.9293	0.9483	0.5092
ShotStream	0.6146	0.1753	0.9617	0.9670	0.9647
Ours	**0.6261**	**0.1980**	**0.9717**	0.9675	**0.9732**

Table 2: Comparison with bidirectional multi-shot generation models under the 15-second setting.

Method	Architecture	Aesthetic ↑	Text Align. ↑	Intra-Shot Cons. ↑		Inter-Shot Cons. ↑	SCA ↑
				Subject	Background		
HoloCine	Bidirectional	0.5842	**0.2050**	0.9728	0.9711	**0.6821**	0.9694
MultiShotMaster	Bidirectional	0.5811	0.2046	0.9626	0.9671	0.6530	0.9678
Ours	Causal, 4step	**0.6194**	0.2004	**0.9823**	**0.9752**	0.6608	**0.9883**

Comparisons. We first compare with autoregressive long-video generation methods, including Self-Forcing [17], Infinity-RoPE [52], LongLive [51], MemFlow [21], and ShotStream [30]. These methods extend generation through causal rollout, KV caching, or long-context positional extrapolation, but most of them are primarily designed for short-context continuation. As shown in Tab. 1 and Fig. 3, they often produce locally smooth videos that remain semantically static, repeating similar layouts or missing requested shot-level changes. Our method achieves the best overall performance, with clear gains in text alignment and shot-cut accuracy, showing stronger ability to follow changing per-shot instructions while preserving subject consistency.

We also compare with bidirectional multi-shot models [31, 43], which generate the full sequence jointly. Note that, to align with the preferred generation length of these bidirectional baselines, we evaluate this comparison under their 15s setting. As shown in Tab. 2 and Fig. 4, our causal generator achieves comparable visual quality and cross-shot coherence, while being substantially faster at inference. In addition, our method naturally supports interactive continuation, where users can append new shot prompts during generation without providing the entire prompt sequence in advance.

4.1 Ablation on Key Design Choices

Ablation on Long Multi-Shot Causal Tuning We ablate the ordering of causal multi-shot learning and step compression. Our full framework first adapts the bidirectional video model into a long-context causal multi-shot model, and then performs ODE initialization and DMD distillation for few-step generation. In the ablated setting, we skip this long multi-shot causal tuning stage and directly perform ODE initialization by aligning the student to trajectories from the original 5s bidirectional model. The following DMD stage, student architecture, and inference procedure are kept the same. Table 3 shows that direct compression from the 5s bidirectional model degrades prompt following, shot-cut control, and long-range consistency. This indicates that step compression

Table 3: Ablation studies on causal tuning and memory design.

Ablation	Method	Aesthetic ↑	Text Align. ↑	Intra-Shot Cons. ↑		Inter-Shot Cons. ↑	SCA ↑
				Subject	Background		
Causal tuning	w/o multi-shot tuning	0.5967	0.1921	0.9311	0.9519	0.5034	0.5042
	w/ multi-shot tuning	**0.6261**	**0.1980**	**0.9717**	**0.9675**	**0.6529**	**0.9732**
Memory	w/o memory	0.5827	0.2181	0.9432	0.9412	0.5832	**0.9772**
	first-frame sink	**0.6017**	0.2285	0.9575	0.9443	0.6106	0.9618
	content routing (ours)	0.5974	**0.2394**	**0.9628**	**0.9529**	**0.7530**	0.9745

cannot reliably recover causal multi-shot behavior when it is absent from the initialization. As shown in Fig. 5, the ablated model suffers from unstable intra-shot content and inconsistent cross-shot identity, while our full pipeline remains more stable.

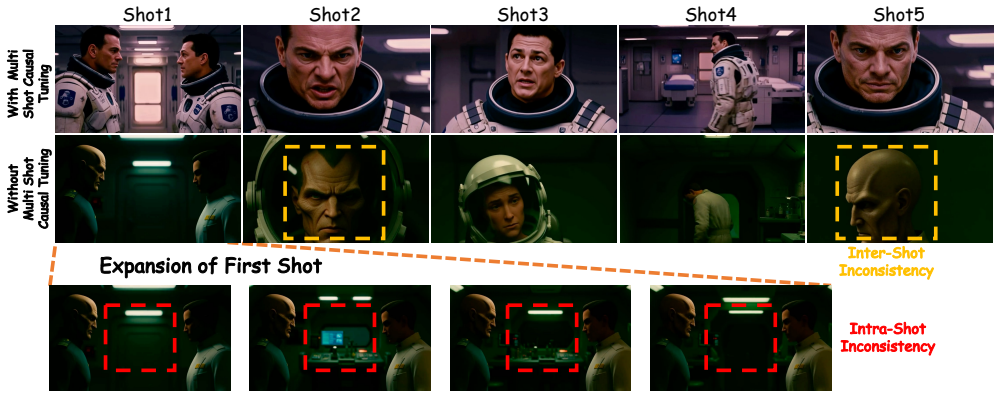


Figure 5: Effect of learning causal multi-shot structure before step compression. Directly initializing a few-step student from a short bidirectional teacher leads to unstable intra-shot content and inconsistent identities across shots. Our full pipeline first learns long-context causal multi-shot generation and then compresses sampling steps, improving temporal stability and cross-shot identity preservation.

Ablation on Memory Design We study how memory design affects long multi-shot rollout. To better evaluate the memory mechanism, we construct a dedicated evaluation set comprising 100 memory test prompts generated by Gemini 2.5 pro [6], which specifically emphasize scenarios where subjects disappear and reappear across shots. Table 3 compares three variants: removing long-term memory, using first-frame sink memory, and our content-aware memory routing. Without long-term memory, the model mainly relies on the local KV window and often forgets entities after long temporal gaps. First-frame sink memory provides a fixed positional anchor, but the earliest frames are not necessarily relevant after several shot cuts. Content-aware memory routing achieves the best inter-shot consistency by retrieving historical frames according to semantic affinity rather than temporal position. As shown in Fig. 6, both the no-memory and first-frame-sink variants fail to faithfully recover the character when it reappears in the final shot. In contrast, our method retrieves the earlier character shot and preserves distinctive identity cues.

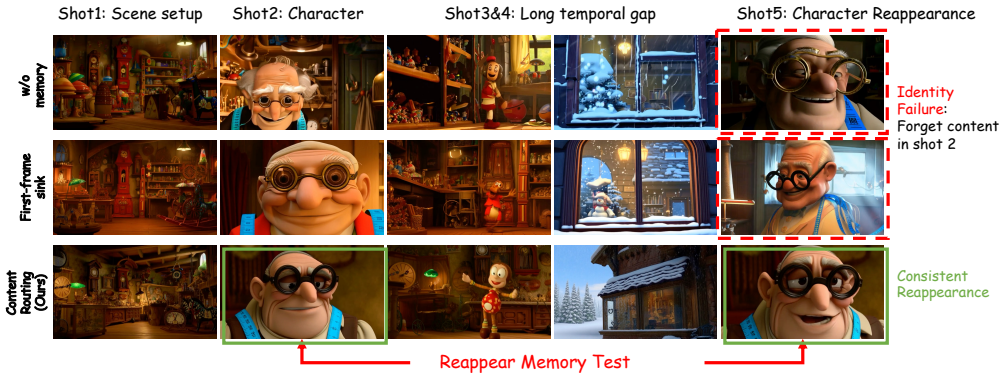


Figure 6: Content-aware memory routing better preserves character identity across long temporal gaps than no-memory and first-frame sink variants, enabling a consistent reappearance in final shot.

5 Conclusion

We presented CausalCine, a causal framework for interactive multi-shot video generation. By learning long-form shot transitions before distillation and routing KV memory by content, CausalCine supports shot-level prompt updates, bounded-attention rollout, and cross-shot recall without regenerating earlier video. Experiments show improved prompt following, shot-cut control, and identity preservation over autoregressive baselines, with visual quality competitive with bidirectional multi-shot models.

Acknowledgements and Disclosures

The authors declare no funding to disclose for this work and no competing interests.

References

[1] 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.

[2] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

[3] Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, et al. Mixture of contexts for long video generation. *arXiv preprint arXiv:2508.21058*, 2025.

[4] Shengqu Cai, Weili Nie, Chao Liu, Julius Berner, Lvmin Zhang, Nanye Ma, Hansheng Chen, Maneesh Agrawala, Leonidas Guibas, Gordon Wetzstein, and Arash Vahdat. Mode seeking meets mean seeking for fast long video generation. In *arXiv*, 2026.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, 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.

[7] Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025.

[8] Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Lol: Longer than longer, scaling video generation to hour. *arXiv preprint arXiv:2601.16914*, 2026.

[9] 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.

[10] Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025.

[11] Yuwei Guo, Ceyuan Yang, Hao He, Yang Zhao, Meng Wei, Zhenheng Yang, Weilin Huang, and Dahua Lin. End-to-end training for autoregressive video diffusion via self-resampling. *arXiv preprint arXiv:2512.15702*, 2025.

[12] Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17281–17291, 2025.

[13] Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

[14] Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2568–2577, 2025.

[15] Yining Hong, Beide Liu, Maxine Wu, Yuanhao Zhai, Kai-Wei Chang, Linjie Li, Kevin Lin, Chung-Ching Lin, Jianfeng Wang, Zhengyuan Yang, et al. Slowfast-vgen: Slow-fast learning for action-driven long video generation. *arXiv preprint arXiv:2410.23277*, 2024.

[16] 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.

[17] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

[18] Yubo Huang, Hailong Guo, Fangtai Wu, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, Jiaming Liu, and Steven Hoi. Live avatar: Streaming real-time audio-driven avatar generation with infinite length, 2025. URL `https://arxiv.org/abs/2512.04677`.

[19] Ziqi Huang, Yinan He, Jiashuo 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.

[20] Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. *arXiv preprint arXiv:2309.14509*, 2023.

[21] Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Memflow: Flowing adaptive memory for consistent and efficient long video narratives. *arXiv preprint arXiv:2512.14699*, 2025.

[22] Weinan Jia, Yuning Lu, Mengqi Huang, Hualiang Wang, Binyuan Huang, Nan Chen, Mu Liu, Jidong Jiang, and Zhendong Mao. Moga: Mixture-of-groups attention for end-to-end long video generation. *arXiv preprint arXiv:2510.18692*, 2025.

[23] Taekyung Ki, Sangwon Jang, Jaehyeong Jo, Jaehong Yoon, and Sung Ju Hwang. Avatar forcing: Real-time interactive head avatar generation for natural conversation. *arXiv preprint arXiv:2601.00664*, 2026.

[24] Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 25690–25699, 2025.

[25] Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*, 2025.

[26] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

[27] Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. Realwonder: Real-time physical action-conditioned video generation. *arXiv preprint arXiv:2603.05449*, 2026.

[28] 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.

[29] Enzhe Lu, Zhejun Jiang, Jingyuan Liu, Yulun Du, Tao Jiang, Chao Hong, Shaowei Liu, Weiran He, Enming Yuan, Yuzhi Wang, et al. Moba: Mixture of block attention for long-context llms. *arXiv preprint arXiv:2502.13189*, 2025.

[30] Yawen Luo, Xiaoyu Shi, Junhao Zhuang, Yutian Chen, Quande Liu, Xintao Wang, Pengfei Wan, and Tianfan Xue. Shotstream: Streaming multi-shot video generation for interactive storytelling. *arXiv preprint arXiv:2603.25746*, 2026.

[31] Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Yixuan Li, Cheng Chen, Yanhong Zeng, et al. Holocine: Holistic generation of cinematic multi-shot long video narratives. *arXiv preprint arXiv:2510.20822*, 2025.

[32] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[33] 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.

[34] Tianhao Qi, Jianlong Yuan, Wanquan Feng, Shancheng Fang, Jiawei Liu, SiYu Zhou, Qian He, Hongtao Xie, and Yongdong Zhang. Maskˆ 2dit: Dual mask-based diffusion transformer for multi-scene long video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18837–18846, 2025.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[37] Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

[38] Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. *arXiv preprint arXiv:2511.01266*, 2025.

[39] Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11218–11221, 2024.

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

[41] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[42] Jiahao Wang, Hualian Sheng, Sijia Cai, Weizhan Zhang, Caixia Yan, Yachuang Feng, Bing Deng, and Jieping Ye. Echoshot: Multi-shot portrait video generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[43] Qinghe Wang, Xiaoyu Shi, Baolu Li, Weikang Bian, Quande Liu, Huchuan Lu, Xintao Wang, Pengfei Wan, Kun Gai, and Xu Jia. Multishotmaster: A controllable multi-shot video generation framework. *arXiv preprint arXiv:2512.03041*, 2025.

[44] Wen Wang, Canyu Zhao, Hao Chen, Zhekai Chen, Kecheng Zheng, and Chunhua Shen. Autostory: Generating diverse storytelling images with minimal human efforts. *International Journal of Computer Vision*, 133(6):3083–3104, 2025.

[45] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

[46] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[47] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

[48] Junfei Xiao, Ceyuan Yang, Lvmin Zhang, Shengqu Cai, Yang Zhao, Yuwei Guo, Gordon Wetzstein, Maneesh Agrawala, Alan Yuille, and Lu Jiang. Captain cinema: Towards short movie generation. In *The Fourteenth International Conference on Learning Representations*, 2025.

[49] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

[50] 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.

[51] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

[52] Hidir Yesiltepe, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. *arXiv preprint arXiv:2511.20649*, 2025.

[53] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024.

[54] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024.

[55] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22963–22974, 2025.

[56] Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11, 2025.

[57] Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *arXiv preprint arXiv:2504.12626*, 2025.

[58] 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.

[59] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

[60] 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: Step-by-step generating multi-shot video with minimal manual intervention. *arXiv preprint arXiv:2412.02259*, 2024.

[61] 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.

[62] Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

A More Results

To provide a more complete view of the generated videos beyond the still frames shown in the main paper, we include additional **video results** in the supplementary material. These examples cover diverse multi-shot prompts, including character reappearance, viewpoint changes, scene transitions, and long-range cross-shot consistency.

We also include a recorded real-time interactive generation demo. The demo shows how CausalCine generates a video chunk by chunk, accepts newly appended shot-level prompts during rollout, and continues the sequence without regenerating previous shots.

For convenient browsing, the supplementary material contains an HTML gallery that organizes all video results and the interactive demo in one place. Readers can open the HTML file directly to view the results without clicking each video file individually.

B Causal Base Model vs. Four-Step Student

We compare the full-step causal base model with the four-step DMD student to evaluate how much quality is retained after acceleration. The causal base is trained by long multi-shot teacher forcing and serves as the high-quality autoregressive teacher for distillation. As shown in Tab. S1 and Fig. S1, it already produces coherent multi-shot rollouts with clear shot transitions, strong prompt following, and stable recurring subjects.

The four-step student preserves these properties while reducing inference cost. Quantitatively, the student remains close to the causal base on visual quality, text alignment, consistency, and shot-cut accuracy. Qualitatively, Fig. S1 shows that the student follows the same shot progression and maintains the main subject across the sequence, despite using only four denoising steps. This indicates that DMD successfully compresses the denoising trajectory while maintaining the multi-shot structure, validating our design of first learning causal long-form behavior in the base model before distillation.



Figure S1: The four-step DMD student preserves the shot-level structure, subject identity, and visual composition of the 50-step causal base while substantially reducing the number of denoising steps.

Table S1: Comparison between the causal base model and the four-step distilled student.

Method	Steps	Aesthetic ↑	Text Align. ↑	Intra-Shot Cons. ↑		Inter-Shot Cons. ↑	SCA ↑
				Subject	Background		
Causal base	50	0.5930	**0.2016**	0.9628	0.9619	**0.6621**	0.9605
DMD student	4	**0.6261**	0.1980	**0.9717**	**0.9675**	0.6529	**0.9732**

C Effect of Adversarial Regularization

We ablate the lightweight adversarial regularization used during DMD distillation by comparing the four-step student trained with and without the GAN head. As shown in Fig. S2, the adversarial regularization stabilizes the sequence-level spatial distribution of long causal rollouts. Without the GAN head, the student still follows the multi-shot prompt, but with noticeable drift in camera motion and subject framing. In particular, recurring subjects may move away from the center of the frame, and some shots exhibit unnatural spatial shifts or composition changes. With GAN regularization, the generated shots maintain plausible camera motion and more stable subject placement.



Figure S2: Ablation of adversarial regularization in DMD distillation. The student trained with the GAN regularization produces more stable subject framing and more plausible camera motion, while the model without GAN regularization shows sequence-level drift and irregular camera motion.

D Limitations and Failure Case

Limitations. CausalCine targets a challenging setting: real-time autoregressive generation of coherent multi-shot videos with prompt changes, shot transitions, and long-range entity recall. To achieve multi-shot fidelity, we use Wan2.1-T2V-14B as the backbone rather than a smaller 1.3B-scale model commonly used by several prior autoregressive video systems, which also increases inference cost. With our distributed deployment, CausalCine reaches real-time generation at 16 FPS on 8 NVIDIA H200 GPUs, but this is still beyond the capability of consumer-grade GPUs. We view this mainly as a systems and model-scaling limitation rather than a fundamental limitation of the causal formulation: future smaller video backbones, quantization, faster attention kernels, and more optimized serving infrastructure could reduce the hardware requirement.

Failure case. A typical failure case is fine-grained physical-state continuity across cuts. CausalCine is designed to preserve high-level narrative context, shot-level prompts, and recurring entities, but it does not explicitly maintain a structured state for small objects, contact geometry, or ongoing physical interactions. As a result, the model can generate individually plausible shots that do not compose into a fully consistent physical process. For example, in the coffee-making failure case in Fig. S3, the scene, cup, and latte-art theme remain recognizable, but the milk stream, pitcher position, hand pose, and foam pattern change in ways that are not physically continuous across cuts. This suggests that content-aware KV memory helps recall visual evidence, but does not by itself solve precise object-state tracking or action-level causality. Future work could combine causal video generation with explicit object-state memory, action constraints, or 3D-aware representations.



Figure S3: Failure case on fine-grained physical-state continuity. CausalCine produces visually plausible coffee-making shots, but the milk stream, pitcher pose, hand position, and latte-art pattern do not evolve as a single physically consistent action across cuts.

E Reproducibility and Experimental Details

E.1 Base Model and Data Format

We build CausalCine on Wan2.1-T2V-14B. The generated videos in our main experiments use resolution 832×480, 481 or 241 frames, and 16 FPS, corresponding to approximately 30 or 15 seconds. We operate in the video VAE latent space. Each autoregressive chunk contains 3 latent frames, corresponding to approximately 12 decoded video frames. The training data consists of 100k long multi-shot videos with shot-level text annotations. Each training sample is stored as a latent file containing the video latent, the global text embedding, per-shot text embeddings, shot

boundary metadata, latent frame count, latent height, and latent width. Shot boundaries are stored in video-frame coordinates and converted to latent-frame boundaries during training and inference. Each chunk is assigned the prompt of the shot that contains it, and both clean context chunks and noisy query chunks use the same shot-indexed text condition.

To make the annotation format concrete, a typical training prompt is organized into a global caption, optional character descriptions, and a sequence of shot-level captions separated by shot-cut markers:

Example training prompt

[global caption] The scene is a magical encounter in a misty, ancient Celtic ruin at dawn. The environment is comprised of mossy standing stones and rolling green hills shrouded in morning mist. This scene contains 5 shots.

[character1] A modern-day historian, a skeptical woman with practical hiking gear and a camera.

[character2] The spectral figure of an ancient Celtic queen, translucent and ethereal, with long, flowing red hair and a silver circlet.

[per-shot captions]

1. Medium shot of [character1] carefully touching a moss-covered standing stone, a look of academic interest on her face.
2. Close-up of her face, her expression changing to one of utter shock as she sees something off-camera.
3. A soft-focus shot of [character2] slowly materializing from the mist between two stones.
4. Medium shot of [character1] stumbling backward, lowering her camera, her skepticism completely shattered.
5. Close-up of [character2]'s spectral face, her expression sad and timeless as she looks at the historian.

The global caption specifies the overall story, setting, character identities, and scene-level context shared by the entire video. The character tags provide reusable entity descriptions that can be referenced by multiple shots, while the per-shot captions specify the local composition, action, and viewpoint for each shot. The `[shot cut]` markers define the temporal segmentation, and these boundaries are converted into latent-frame indices so that each autoregressive chunk is conditioned on the corresponding shot caption during training and inference.

E.2 Causal Multi-Shot Tuning

The causal base model is trained with the parallel teacher-forcing objective described in Sec. 3.2. For a video with N chunks, we pack N clean chunks and N noisy chunks into one $2N$-segment forward pass. Clean chunks use timestep 0, noisy chunks share a sampled diffusion timestep, and the loss is computed only on the noisy half. The main causal tuning configuration is:

- Base model: Wan2.1-T2V-14B.
- Resolution and length: 832×480, 241 frames.
- Chunk size: 3 latent frames.
- Flow-matching schedule shift: 5.0.
- Optimizer: AdamW.
- Learning rate: 1×10^{-5}.
- Warmup steps: 0.
- Gradient clipping: max norm 10.0.
- Classifier-free guidance dropout during training: 0.0.
- Parallelism: FSDP with context-parallel attention.

For CAMR, we use a local context window of $W=3$ chunks and retrieve $k=5$ historical latent frames outside the local window. For long-rollout inference with CAMR, we use Block-Relative RoPE so

that retrieved memory frames, local-window frames, and current frames are re-indexed within the training temporal range.

E.3 ODE Trajectory Generation and Initialization

Before DMD training, we initialize the four-step student with causal ODE distillation. We first generate teacher trajectories from the causal base model using a 48-step PF-ODE solver. The trajectory generation uses:

- Number of teacher trajectory steps: 48.
- Trajectory guidance scale: 6.0.
- Flow shift: 5.0.
- Chunk size: 3 latent frames.
- Evaluation/inference timesteps for the four-step student: $\{0, 250, 500, 750\}$.
- Zero-noise clean context for teacher-forced history.

For each sample, we store intermediate trajectory states used by the ODE initialization stage. The four-step student is trained by randomly sampling one of the four selected ODE states and predicting the teacher's final denoised target under the same causal visibility pattern. ODE initialization uses AdamW with learning rate 2×10^{-6}, max gradient norm 10.0, 2000 optimizer steps, and checkpoint interval 200.

E.4 DMD Distillation with Adversarial Regularization

After ODE initialization, we refine the student using Distribution Matching Distillation (DMD) with adversarial regularization. The generator is initialized from the ODE-initialized causal student checkpoint, while the fake denoiser and frozen real denoiser are initialized from the causal multi-shot training pipeline. The generator performs causal rollout with KV caching and the same per-shot text routing mechanism used at inference. The DMD configuration is:

- Number of student denoising steps: 4.
- Student timesteps: $\{0, 250, 500, 750\}$.
- Generator update ratio: 5.
- Fake denoiser update ratio: 1.
- Generator optimizer: AdamW with learning rate 2×10^{-6}, β_1=0.0, β_2=0.999.
- Fake denoiser optimizer: AdamW with learning rate 4×10^{-7}, β_1=0.0, β_2=0.999.
- Training length: 1000 optimizer steps.
- Classifier-free guidance scale: 1.0, since guidance is distilled into the student trajectory.

The adversarial regularization uses a lightweight discriminator head trained jointly with DMD:

- Discriminator update ratio: 5.
- Adversarial loss weight: 0.1.
- Maximum discriminator timestep: normalized timestep 0.3.
- Discriminator head optimizer: AdamW with learning rate 5×10^{-7}, β_1=0.0, β_2=0.999.
- Discriminator feature layers: layers 29 and 39.

E.5 Computation Resource

The main experiments are conducted on a distributed GPU cluster. Each worker node contains 8 NVIDIA H800 GPUs with approximately 80GB GRAM per GPU. We use PyTorch distributed training with FSDP and context-parallel attention.

The main compute budget is summarized below:

- Causal multi-shot tuning: 64 GPUs, approximately 3 days.
- ODE trajectory generation: 64 GPUs, approximately 1 days.
- ODE initialization: 64 GPUs, approximately 10 hours.
- DMD distillation: 64 GPUs, approximately 10 hours.

The total research project required more compute than the final reported training runs because it included preliminary experiments, debugging runs, memory-design ablations, causal-tuning ablations, and intermediate checkpoint evaluations. The reported compute above corresponds to the main reproducible experimental pipeline rather than all exploratory runs.

F Broader Impact

CausalCine aims to make long-form video generation more efficient, interactive, and controllable. Potential positive impacts include lowering the cost of creative video production, enabling interactive storyboarding, supporting rapid prototyping for filmmakers and designers, and reducing the need to regenerate an entire video when only future shots need to be extended or modified. The causal formulation can also reduce inference cost by reusing KV caches and avoiding recomputation of previously generated content. At the same time, improved video generation technology can be misused. Higher-quality and faster video synthesis may facilitate deceptive synthetic media, impersonation, misinformation, spam, or copyright-infringing content generation. Multi-shot generation may further increase the persuasive quality of generated narratives by allowing coherent story progression across shots. The model may also reproduce biases present in its training data, including stereotypes about people, places, professions, or cultural contexts. Since the system generates visual content from text prompts, harmful or ambiguous prompts can lead to unsafe or misleading outputs if not filtered by a deployment system. This paper studies the modeling and efficiency aspects of causal multi-shot generation and does not propose a complete deployment safety stack. We therefore recommend that any public-facing deployment use prompt filtering, output moderation, visible or invisible watermarking, provenance metadata, abuse monitoring, and clear disclosure that generated videos are synthetic. We also recommend additional domain-specific evaluation before using the model in sensitive applications such as news, political communication, biometric identity, surveillance, medical content, or legal evidence.

G Responsible Release and Safeguards

Video demonstrations are available on the project page. Code, model checkpoints, and training data are not publicly released with this paper. Future release plans remain undecided. For downstream applications, we recommend combining CausalCine with safety mechanisms that are outside the scope of this paper, including prompt safety classifiers, generated-content watermarking, content provenance metadata, human review for high-risk use cases, and deployment-time monitoring. These safeguards are particularly important because faster autoregressive generation lowers the cost of producing long synthetic videos.

H Existing Assets

We use Wan2.1-T2V-14B as the pretrained base video diffusion model and cite the corresponding paper. The Wan2.1-T2V-14B model is released under the Apache 2.0 license. Our implementation builds on PyTorch, Hugging Face Diffusers-style model components, and distributed training utilities. For evaluation, we use several existing models and tools. We use VBench-style metrics for video generation evaluation; VBench is released under the Apache 2.0 license. We use CLIP/OpenCLIP-style image-text features, which are released under MIT-style licenses, and DINO/DINOv2 features for visual consistency evaluation; DINOv2 code is released under Apache 2.0. We use the LAION aesthetic predictor, which is released under the MIT license, and TransNetV2 for shot-boundary detection, which is released under the MIT license. We use ViCLIP text-video similarity for prompt following. We use Gemini 2.5 Pro to help construct the 100-prompt multi-shot evaluation benchmark. Gemini is not part of the CausalCine model, training objective, or inference algorithm. Its use is limited to generating and organizing evaluation prompts containing global story descriptions, shot-level captions, and target shot boundaries.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scope as causal, interactive multi-shot video generation with long multi-shot causal tuning, content-aware memory routing, and few-step distillation. These claims are supported by the method in Sec. 3 and the comparisons and ablations in Sec. 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations and failure cases in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theoretical results. The equations in Sec. 3 define the training objectives, causal factorization, and memory routing procedure, but we do not claim new theorems requiring proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No]

Justification: We provide the model architecture, training stages, data format, evaluation protocol, metrics, inference settings, and compute details in Sec.4 and AppendixE. Access to the internal training dataset, training code, model checkpoints, and inference/evaluation code is currently restricted. These access limitations constrain independent reproduction of the reported results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code, model checkpoints, and training data are not publicly released with this paper. Video demonstrations are available at `https://yihao-meng.github.io/CausalCine/`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer:[Yes]

   Justification: The main paper specifies the base model, resolution, chunk size, training data scale, evaluation benchmark, and metrics; Appendix E provides the full training and inference hyperparameters. These include optimizer settings, denoising steps, context window, CAMR top-k, shot-cut handling, and evaluation protocol

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Answer: [No]

   Justification: We evaluate all methods on the same fixed 100-prompt multi-shot benchmark and report deterministic metric averages, but we do not report error bars or confidence intervals. Running multiple seeds for all 14B-scale baselines (especially bidirectional baselines) would be prohibitively expensive.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E.5 reports the GPU type, number of nodes, GPUs per node, memory per GPU, and training/evaluation configuration for the main experiments. We also disclose that the full research process involved additional preliminary and ablation runs beyond the final reported checkpoints.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and believe the research conforms to it. The work uses licensed existing assets, does not involve human-subject experiments, and discusses risks and safeguards for generative video release in the appendix.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix F discusses positive impacts such as more efficient interactive video creation and negative risks such as deceptive synthetic media, copyright concerns, and misuse for disinformation. We also discuss release restrictions and documentation as mitigation.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: Appendix G discusses safeguards recommended for potential future deployment. The paper does not release model checkpoints or training data, and does not claim to have implemented a complete deployment safety system.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the existing models, datasets, and evaluation tools used in the paper and summarize their licenses or terms in Appendix H. These include Wan2.1, internal multi-shot video datasets with license, VBench-style metrics, CLIP/DINO-family evaluators, TransNetV2, LAION aesthetic predictor, and Gemini API usage for prompt construction.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper presents a new model and experiments, but does not release code, model checkpoints, or datasets. The public project page provides video demonstrations.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects. No participant instructions, screenshots, or compensation details are applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subject research or crowdsourcing studies, so IRB or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core CausalCine model, training objective, and inference algorithm do not use an LLM. Gemini 2.5 Pro was used to construct evaluation prompts, as described in Sec. 4 and Appendix H.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.