CONTEXT-MATCHED DISTILLATION: TEACHER CAUSALITY FOR AUTOREGRESSIVE VIDEO DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive autoregressive video generation demands both low-latency rollouts and precise online control. Few-step distillation accelerates generation by reducing denoising steps, while online control imposes a causal constraint: frames and blocks should depend on history and controls available during generation. Existing video distribution matching distillation (DMD) pipelines, however, often supervise causal few-step students using bidirectional teachers that score complete clips. The score for a target can therefore depend on future frames and controls that were unavailable when the student generated it, misaligning teacher supervision with the student's causal information set. We introduce **Context-Matched Distillation (CMD)**, a causal DMD framework that aligns teacher supervision with the information available when each target is generated. CMD replaces bidirectional full-clip scoring with a causal teacher that evaluates each target without access to future frames or controls. The same causal teacher initializes the few-step student, establishing a consistent causal formulation across teacher training, student distillation, and inference. Beyond aligning the temporal information boundary, **Prefix Scoring** matches supervision to the student's realized rollout context by evaluating each target under the cached student-generated prefix that produced it. **Prefix Corruption** further stabilizes training by perturbing unreliable prefixes produced early in training while preserving this target–context alignment. With a simple causal formulation, CMD naturally extends to frame-wise and chunk-wise generation, long video distillation, and camera-conditioned distillation. Experiments demonstrate state-of-the-art aggregate performance among autoregressive methods on both short- and long-video benchmarks, together with substantially improved adherence to time-varying camera controls.

1 INTRODUCTION

Autoregressive (AR) video models generate visual observations sequentially, making them a natural fit for interactive generative world models (Yin et al., 2025; Huang et al., 2025a; Team et al., 2026; Zhu et al., 2026a). To support practical interaction, these models must sustain long-horizon rollouts, accept frequent control updates, and faithfully execute time-varying control signals such as camera poses. Every generated frame or block is reused as context for the next prediction, so even small errors can compound into long-term drift in appearance, geometry, and motion.

Few-step distillation reduces the cost of each autoregressive transition by compressing multi-step denoising into a small number of student evaluations (Yin et al., 2024b;a; 2025). Methods such as Self-Forcing further roll out the student during training, exposing it to its own generated histories and reducing exposure bias (Huang et al., 2025a; Guo et al., 2025). However, standard distribution matching distillation (DMD) pipelines still use a bidirectional teacher to score each completed generated clip. A target can therefore be supervised using future frames and controls that were unavailable when the causal student produced it. We call this the *teacher–student context mismatch*.

The context mismatch is especially consequential for camera control and long video generation. In camera-conditioned distillation, a bidirectional teacher sees the complete camera trajectory, while the student must execute each pose update online (He et al., 2024; Wang et al., 2024). Future pose changes can therefore affect the supervision assigned to the current transition. For long rollouts, treating each local clip as self-contained both discards the preceding rollout context used to generate its early targets and allows those targets to attend to later frames inside the scoring window (Cai et al., 2026; Yang et al., 2025). In both cases, the teacher scores a target under a different context from the one in which the deployed generator acts.

To align these contexts, we introduce **Context-Matched Distillation (CMD)**, which scores each target with a causal teacher that cannot access later frames or controls. CMD first fine-tunes a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

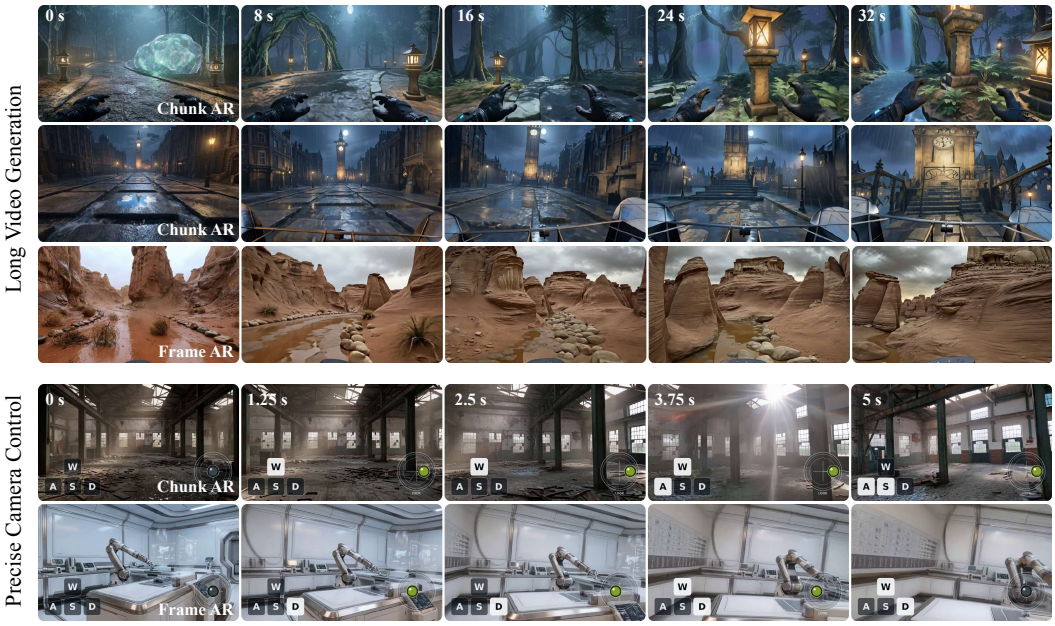


Figure 1: **Context-Matched Distillation supports long-horizon and camera-controllable autoregressive video generation from a single image.** Starting from the conditioning image at 0 s (left), our frame-wise (Frame AR) and multi-frame (Chunk AR) models extend each scene. **See the supplementary video gallery for short-video, long-video, and camera-controlled results from both chunk-1 and chunk-4 models.**

pretrained bidirectional model into a multi-step causal teacher with a Diffusion-Forcing objective over independently noised histories (Chen et al., 2024), then initializes the few-step student from those weights. During distillation, the student is rolled out autoregressively, while the frozen teacher scores its noised targets causally. The same causal architecture and conditioning are therefore used throughout teacher training, distillation, and student inference. As a practical side benefit, adding a new control does not require separately adapting both a bidirectional conditional teacher and a causal student, and no additional ODE-matching or consistency-distillation stage is needed for initialization (Yin et al., 2025; Zhu et al., 2026b; Zhao et al., 2026; Gao et al., 2026).

Causal scoring prevents future information from affecting a target score, but it does not ensure that the teacher observes the true history that produced the target. A naive implementation conditions on preceding noised DMD targets rather than the realized student rollout. We therefore introduce **Prefix Scoring**, which supplies the cached student-generated prefix used to produce each target. Early student prefixes may contain structured rollout artifacts unlike the independent corruptions seen during teacher training. **Prefix Corruption** therefore perturbs the cached prefix to reduce the teacher's sensitivity to these unreliable details while preserving causal ordering and control signals.

CMD's causal target–prefix scoring applies naturally to frame-wise and chunk-wise generation, bounded-context long videos, and time-varying camera control. CMD consistently improves short- and long-video quality over bidirectional scoring baselines. For camera-conditioned generation, it substantially reduces camera-pose errors, confirming the importance of matching teacher supervision to temporal information from online generation. Ablations further isolate the contributions of causal scoring, matched rollout prefixes, and prefix corruption. **We discuss related work and provide a broader review in Supplementary Section S1.** Our contributions are:

- We introduce **Context-Matched Distillation**, which uses a causal teacher to score every generated target without access to later frames or controls.
- We introduce **Prefix Scoring** to condition each score on the student-generated history that produced its target, together with **Prefix Corruption** for robust training on unreliable early rollouts.
- We apply the same formulation across frame-wise and chunk-wise generation, long-video rollout, and camera control, achieving strong short- and long-video quality and substantially lower camera-pose errors.

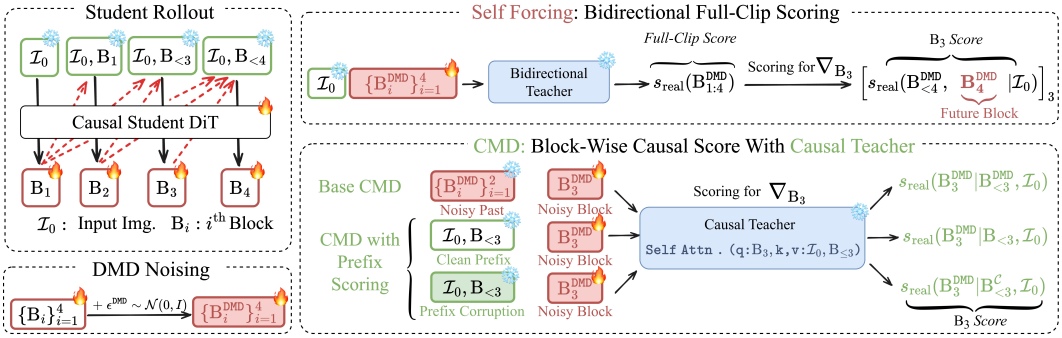


Figure 2: **Context-Matched Distillation.** A causal student first generates an autoregressive rollout. Self-Forcing scores the full noised rollout with a bidirectional teacher, allowing the score of a target to depend on future blocks. CMD instead uses a causal teacher: Base CMD conditions on preceding noised DMD blocks, Prefix Scoring uses the cached clean student prefix that produced the target, and Prefix Corruption perturbs this prefix to stabilize supervision when early rollouts are unreliable.

2 PRELIMINARIES

Distribution Matching Distillation. Distribution matching distillation (Yin et al., 2024b;a) distills a few-step student generator G_θ from a multi-step teacher by matching the student distribution p_{fake} to the teacher distribution p_{real}. DMD uses the difference between their scores, s_{fake} and s_{real}, as a proxy for the distribution mismatch and minimizes this. The resulting gradient is written as

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = -\mathbb{E}_{x\sim p_{\text{fake}}}\left[\left(s_{\text{real}}(x) - s_{\text{fake}}(x)\right)\frac{\mathrm{d}G_\theta}{\mathrm{d}\theta}\right]. \tag{1}$$

For readability, we suppress the diffusion timestep and write $s(x)$ for the score evaluated on a noised version of x, *i.e.* $x_\tau = (1-\tau)x + \tau\cdot\epsilon$ where $\epsilon\sim\mathcal{N}(0,I)$, and timestep is $\tau\in[0,1]$.

While DMD was introduced for distillation of few-step text-to-image models (Yin et al., 2024b;a; Bandyopadhyay et al., 2026), it has recently been applied to distilling text-to-video and image-to-video models, including causal and frame-wise generators (Yin et al., 2025; Zhang et al., 2025b; Huang et al., 2025a; Zhu et al., 2026b; Zhao et al., 2026; Zheng et al., 2026; Feng et al., 2026). In these applications, the distillation strategy largely follows the T2I setting: the student generates a full sample, and the teacher provides a score over the corresponding full video clip.

Autoregressive Video Generation. We represent an image-to-video latent sequence as $x \equiv \mathbf{x}_{0:T} = (\mathbf{x}_0,\ldots,\mathbf{x}_T)$, where $\mathbf{x}_0 = \mathcal{I}_0$ is the fixed input-image latent and $\mathbf{x}_{1:T}$ are the generated video latents. The sequence $c_{1:T}$ represents per-frame conditioning signals such as camera motion or user controls for the generated frames. Standard video diffusion models (Diffusion Transformers) generate a full video clip jointly, typically using bidirectional temporal attention over all frames. This allows each latent frame to depend on both past and future frames during denoising, but the cost of full attention grows quadratically with the number of frames. As a result, extending such models to long or interactive generation is expensive. Autoregressive video generation instead decomposes the video distribution over time (Gao et al., 2024; Yin et al., 2025; Huang et al., 2025a). In a chunk-wise autoregressive model, we partition the generated latents and their controls into N aligned blocks $\mathrm{B}_{1:N}$ and $C_{1:N}$, each containing K frames. The model factorizes as

$$p_\theta(\mathrm{B}_{1:N} \mid \mathcal{I}_0, C_{1:N}) = \prod_{i=1}^{N} p_\theta(\mathrm{B}_i \mid \mathcal{I}_0, \mathrm{B}_{<i}, C_{\le i}). \tag{2}$$

Causality is enforced across generated blocks, while frames within the current block may interact bidirectionally. We treat the input image $\mathcal{I}_0$ as part of the ordinary causal history rather than something that is permanently retained beyond local attention windows. Setting $K=1$ recovers frame-wise autoregressive generation (Zhao et al., 2026; Feng et al., 2026), whose factorization is

$$p_\theta(\mathbf{x}_{1:T} \mid \mathbf{x}_0 = \mathcal{I}_0, c_{1:T}) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_t \mid \mathbf{x}_{<t}, c_{\le t}). \tag{3}$$

Diffusion Forcing. Diffusion Forcing (Chen et al., 2024) trains causal sequence models by independently corrupting frames and denoising each target under a noisy history. By varying the corruption applied to past frames, it exposes the model to imperfect context during training rather than conditioning only on clean data, improving robustness when generated frames are context during inference. We use this objective to pre-train the causal teacher used for Context-Matched Distillation.

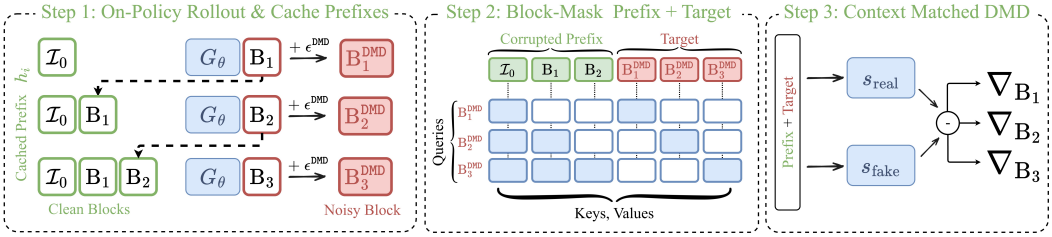


Figure 3: **Efficient Prefix Scoring.** The causal student first generates an on-policy rollout and caches the clean prefix used to produce each block. Corrupted prefixes and their corresponding noised DMD targets are then packed under a block-causal attention mask, allowing all targets to be scored in parallel. Finally, the real and fake scores are evaluated under the same prefix–target context to form the context-matched DMD update.

3 METHODOLOGY

We introduce **Context-Matched Distillation (CMD)** to distill a multi-step causal teacher into a few-step causal student for autoregressive video generation. We begin by fine-tuning a pretrained bidirectional teacher into multi-step causal and block-causal variants. The student models are initialized from the teacher weights and trained by scoring their generated samples against the teacher distribution. Specifically, the student is rolled out on-policy to produce generated targets, while the causal teacher scores each target under the same temporal information boundary as the student, without access to future frames or controls. Building on base CMD, **Prefix Scoring** replaces the noised scoring history for the current frame with the cached student-generated prefix used to generate it. A block-causal attention mask pairs every target with its corresponding prefix while allowing all targets to be scored in a single teacher pass. We further introduce **Prefix Corruption** to reduce sensitivity to artifacts in a weak student's generated prefixes. Section 3.1 motivates CMD by identifying the information-set mismatch between bidirectional teacher scoring and causal student generation. Section 3.2 presents the core distillation method, including causal-teacher pre-training, Prefix Scoring, efficient block-causal scoring, and Prefix Corruption. Finally, section 3.3 and section 3.4 naturally extend CMD to long-video generation and camera control with minimal changes to its formulation.

3.1 MOTIVATION

Conditioning mismatch in causal training. We discuss the formulation for frame-wise causal video generation pipelines in eq. (3). This factorization makes direct DMD supervision from a bidirectional teacher problematic (Zhu et al., 2026b), where $s_{\text{real}}(\mathbf{x}_{1:T}) = \nabla_{\mathbf{x}_{1:T}} \log p_{\text{real}}(\mathbf{x}_{1:T} \mid \mathbf{x}_0 = \mathcal{I}_0, c_{1:T})$. The fake sample in eq. (1) is $\mathbf{x}_{1:T}$, but each frame $\mathbf{x}_t$, $t = 1, \ldots, T$, is generated using only the causal information set $(\mathbf{x}_{<t}, c_{\leq t})$. In contrast, a bidirectional teacher provides $s_{\text{real}}(\mathbf{x}_{1:T})$ by scoring the full video in its native full-clip denoising setting. Its supervision for frame t can therefore depend on information that is unavailable when the causal student generates that frame, including future latents $\mathbf{x}_{>t}$ and future conditioning $c_{>t}$. Thus, although the DMD update is applied to a student sample, the teacher score is computed under a different information set from the one used by the student at inference time.

Context-Matched Distillation (CMD) addresses this conditioning mismatch by replacing the bidirectional full-clip teacher score with that from a causal teacher queried under the same information set available to the student. For $t = 1, \ldots, T$, we use the causal information set $h_t = (\mathbf{x}_{<t}, c_{\leq t})$, where $\mathbf{x}_{<t}$ corresponds to past frames, including image conditioning $\mathbf{x}_0 = \mathcal{I}_0$. The teacher score used in DMD for frame t can be written as $s_{\text{real},t}(\mathbf{x}_t, h_t) = \nabla_{\mathbf{x}_t} \log p_{\text{real}}(\mathbf{x}_t \mid h_t)$ rather than as the t-th component of a bidirectional full-clip score (see fig. 2). The student score is evaluated under the same causal information boundary, yielding a frame-level score difference $s_{\text{real},t}(\mathbf{x}_t, h_t) - s_{\text{fake},t}(\mathbf{x}_t, h_t)$. This preserves the standard DMD score-difference update form while aligning the teacher supervision with the causal information available during autoregressive inference (see fig. 3).

3.2 CONTEXT-MATCHED DISTILLATION

Pre-training Causal Teacher. As discussed in section 3.1, CMD requires teacher scoring to follow the same causal information boundary as student generation. Towards this, we pre-train a causal multi-step teacher with a standard flow-matching objective to perform generation without bidirectional context. Given a real video latent sequence $\mathbf{x}_{0:T}$ with $\mathbf{x}_0 = \mathcal{I}_0$, we keep the input image clean and independently corrupt each input frame with noise at random timesteps. A block mask restricts target frame t to $h_t = (\mathcal{I}_0, \tilde{\mathbf{x}}_{1:t-1}, c_{1:t})$, where $\tilde{\mathbf{x}}_{1:t-1}$ denotes preceding frames corrupted

independently. The teacher is trained with the standard flow-matching loss

$$\mathcal{L}_{\text{DF}} = \mathbb{E}\left\|v_t - \eta_\phi(\mathbf{x}_{t,\tau}, \tau, h_t)\right\|_2^2, \tag{4}$$

where $v_t = (\epsilon - \mathbf{x}_t)$. Following Diffusion Forcing (Chen et al., 2024), this improves robustness to noisy context from the past. The learned flow field $\eta_\phi(\mathbf{x}_{t,\tau}, \tau, h_t)$ can be used to score the causal student pipeline as $s_{\text{real}}(\mathbf{x}_{t,\tau}, \tau, h_t)$, following Yin et al. (2024b).

Causal Student Distillation. We use the pretrained causal teacher η_ϕ for initializing the few-step student G_θ to carry forward the teacher's causal inductive bias. **Notably, our formulation does not require expensive student initialization through ODE (Zhuang et al., 2026; Yin et al., 2025) or Consistency Distillation (Zhu et al., 2026b; Zhao et al., 2026) based pre-training.** During distillation, the student is rolled out exactly as it would be during inference, using the backward-simulation procedure of Yin et al. (2024a), as adopted in Self-Forcing (Huang et al., 2025a). Specifically, we set $\hat{\mathbf{x}}_0 = \mathcal{I}_0$ and use the student G_θ to generate $\hat{\mathbf{x}}_{1:T}$. For $t = 1, \ldots, T$, latent frames are generated with velocities $\hat{v}_{t,\tau} = G_\theta(\hat{\mathbf{x}}_{t,\tau}, \tau, h_t)$, where h_t corresponds to past-frame latents detached from gradient computation. For base CMD, each student-generated target $\hat{\mathbf{x}}_t$ is noised to obtain $\hat{\mathbf{x}}^{\text{DMD}}_{t,\tau}$. These noised targets are arranged temporally and scored with the causal teacher and fake-score network using a block-causal attention mask. The scoring context for target t is $h_t^{\text{CMD}} = (\mathcal{I}_0, \{\hat{\mathbf{x}}^{\text{DMD}}_{i,\tau}\}_{i=1}^{t-1}, c_{1:t})$, where $\hat{\mathbf{x}}^{\text{DMD}}_{<t,\tau}$ contains the preceding noised DMD targets. The causal attention mask prevents the score for $\hat{\mathbf{x}}^{\text{DMD}}_{t,\tau}$ from accessing future frames or controls. Base CMD provides causal teacher supervision matched to the temporal information boundary of the student.

Prefix Scoring. Although base CMD matches the temporal information boundary, its scoring history consists of preceding noised DMD targets rather than the realized history for the student to generate current target. In practice, we instantiate CMD with Prefix Scoring. During the on-policy student rollout, we cache the generated history $\hat{\mathbf{x}}_{<t}$ and controls $c_{\leq t}$ that condition each target. For DMD supervision, frame $\hat{\mathbf{x}}_t$ generated by the student is noised at timestep τ with ϵ^{DMD} to obtain $\hat{\mathbf{x}}^{\text{DMD}}_{t,\tau}$. The frozen causal teacher η_ϕ and the fake-score network evaluate this noised target using the cached causal history $h_t = (\mathcal{I}_0, \hat{\mathbf{x}}_{1:t-1}, c_{1:t})$, yielding $s_{\text{real}}(\hat{\mathbf{x}}^{\text{DMD}}_{t,\tau}, \tau, h_t)$ and $s_{\text{fake}}(\hat{\mathbf{x}}^{\text{DMD}}_{t,\tau}, \tau, h_t)$, respectively, for the update in eq. (1). Rather than using previous noised DMD targets as causal context, Prefix Scoring therefore matches the scoring context to that under which the target was generated. Naively, Prefix Scoring would require a separate teacher pass for each frame. Instead, the teacher receives the input image and student-generated frames as conditioning memory, and uses a block-causal mask which lets each scoring query attend only to its causal prefix $\hat{\mathbf{x}}_{<t}$ and controls $c_{\leq t}$, allowing all generated frames to be scored jointly in one pass, akin to Teacher Forcing (Gao et al., 2024).

Prefix Corruption. Autoregressive video methods have addressed rollout error through self-generated histories (Huang et al., 2025a), self-resampling that simulates imperfect histories (Guo et al., 2025), anti-drifting context construction (Zhang et al., 2025a), and gradients through historical context representations (Zhuang et al., 2026). Complementary to these, a naive teacher scoring on the exact student-generated prefix is not ideal, as the student can drift heavily right after initialization in few-step generation. This can dominate training as causal teacher score is conditioned on the heavily drifted, out-of-distribution prefix. To improve teacher score quality we thus introduce prefix corruption, where the conditioning prefix is corrupted with Gaussian noise to reduce teacher focus on common drift artifacts such as color or texture. Specifically, we corrupt the prefix as:

$$\tilde{h}_t^{(\rho)} = (\mathcal{I}_0, \mathcal{C}_\rho(\hat{\mathbf{x}}_{1:t-1}), c_{1:t}), \qquad \mathcal{C}_\rho(\hat{\mathbf{x}}_{1:t-1}) = (1-\rho)\hat{\mathbf{x}}_{1:t-1} + \rho \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{5}$$

Here ρ is a corruption hyperparameter and is set to a small value by default. The Prefix Scoring DMD objective can now be written as:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = -\mathbb{E}_{\hat{\mathbf{x}}^{\text{DMD}}_{t,\tau}, \tau}\left[\left(s_{\text{real}}(\hat{\mathbf{x}}^{\text{DMD}}_{t,\tau}, \tau, \tilde{h}_t^{(\rho)}) - s_{\text{fake}}(\hat{\mathbf{x}}^{\text{DMD}}_{t,\tau}, \tau, \tilde{h}_t^{(\rho)})\right)\frac{\mathrm{d}G_\theta}{\mathrm{d}\theta}\right]. \tag{6}$$

Separate from noising in naive causal DMD, Prefix Corruption is a controlled relaxation of Prefix Scoring to stabilize distillation, and perturbs the on-policy prefix rather than the base CMD scoring sequence. The prefix is additionally corrupted using a dedicated noise level, independent of randomized DMD timesteps for scoring.

3.3 Long-Video Prefixes

Causal long-video models manage their history through rolling windows and sink frames (Yang et al., 2025; Liu et al., 2025; Yesiltepe et al., 2026), context packing or cache compression (Zhang

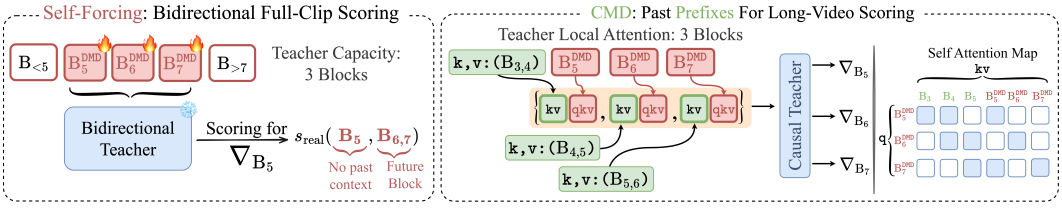


Figure 4: **Prefix Scoring for long-video distillation.** A fixed-capacity bidirectional teacher treats each local clip as self-contained, omitting earlier rollout context while allowing access to future blocks. Prefix Scoring instead evaluates all targets in one causal teacher pass under their corresponding bounded rollout prefixes.

et al., 2025a; Yi et al., 2025; Mao et al., 2026), or learned long-term memory (Yu et al., 2025; Xiao et al., 2025). We use a local attention window for both the student and the causal teacher, as illustrated in fig. 4. Let M denote the number of preceding frames retained by the student. For $t > M$, the student generates frame t under the bounded causal context $h_t^M = (\hat{\mathbf{x}}_{t-M:t-1}, c_{t-M:t})$ in $G_\theta(\hat{\mathbf{x}}_{t,\tau}, \tau, h_t^M)$. Under the same memory budget, a bidirectional teacher instead scores a bounded clip such as $\hat{\mathbf{x}}_{t-M:t}$ as a self-contained sample. This omits the pre-clip context that influenced the clip's earliest frames during the student rollout and allows those frames to attend to later frames in the clip. The resulting score therefore does not match the student's causal factorization.

In contrast, CMD replaces the bidirectional local-clip score with a causal teacher that prevents each target from accessing later frames in the scoring window. With Prefix Scoring, the causal teacher additionally evaluates each target under exactly the same bounded context as the student, *i.e.* $h_t^M = (\hat{\mathbf{x}}_{t-M:t-1}, c_{t-M:t})$. In practice, we train $(RT+1)$-frame videos using R contiguous rollouts of T generated frames, with $\mathcal{I}_0$ included only in the first rollout. After each rollout, its generated frames are cached as context for the next. Consequently, the context of a frame near a rollout boundary may include frames from the preceding rollout. These cached frames provide the local attention context and, when applicable, the sink tokens for both the student and the causal teacher.

For long-video training, we replace the fixed prefix-corruption level ρ in eq. (5) with a frame-dependent schedule ρ_t over the full rollout. Early frames, before substantial drift has accumulated, receive weak prefix corruption, whereas later frames receive stronger corruption as rollout error grows. Applying the score-distillation loss to every generated frame across the R rollouts therefore supervises the full causal trajectory while requiring attention only over a bounded prefix for each target. Longer target videos add rollout chunks to be scored without expanding the teacher's attention window, allowing the same local causal formulation to scale to arbitrarily long videos.

3.4 Camera-Conditioned Causal Distillation

Existing video models encode pose sequences through ray maps, camera adapters, or pose-aware attention (He et al., 2024; Wang et al., 2024; Li et al., 2025; He et al., 2025), and recent autoregressive systems support interactive camera updates during streaming generation (Xu et al., 2026; Team et al., 2026). Prior work has used cameras relative to the previous frame (Team et al., 2026), while relative camera representations have shown improvements in novel-view synthesis quality over absolute alternatives (Li et al., 2025). We adopt this parameterization of frame-relative cameras for training our teacher and student pipelines, embedding the camera as the motion of the current view relative to a generated view. Let $E_t \in SE(3)$ denote the camera-to-world pose for frame $t = 0, \ldots, T$ and K_t the corresponding intrinsics. We convert the trajectory into frame-relative increments:

$$\Delta E_t = \begin{cases} I, & t = 0, \\ E_{t-1}^{-1} E_t, & t > 0, \end{cases} \tag{7}$$

and use them, rather than the absolute pose for camera condition. For each latent-frame location (u, v), we unproject rays using K_t and transform them using ΔE_t for spatial ray embedding

$$c_t = \Phi_t(u, v) = \phi(\Delta E_t, K_t, u, v), \tag{8}$$

where ϕ is implemented as a ray map (Team et al., 2026). This representation preserves the local geometry of the commanded camera motion, while removing dependence on an arbitrary global coordinate frame. The key distinction is how the camera representation is used by the teacher score. In full-clip distillation, even a relative camera trajectory can be provided to a bidirectional teacher as a complete sequence. The score for frame t may then depend on future camera updates $\Phi_{>t}$, allowing the teacher to explain the current frame using motion that is unavailable to the autoregressive student at inference time. This creates the same information-set mismatch as future-frame lookahead:

Table 1: SANA-WM camera benchmark (Simple/Hard mean): Q, S, Total, Rot., Trans., CamMC.

Method	Chunk Size	VBench Scores ↑			Camera Error ↓		
		Quality	Semantic	Total	Rot.	Trans.	CamMC
HY-WorldPlay	4	0.8197	0.1763	0.6910	4.5916	0.2061	0.2546
LingBot-World	3	0.8159	0.1817	0.6890	5.4707	0.1668	0.2406
SANA-WM	3	**0.8309**	0.1067	0.6861	1.6279	0.1569	0.1699
minWM	4	0.8135	0.1801	0.6868	4.8883	0.2493	0.2963
Ours	1	0.8222	**0.2418**	**0.7062**	1.9641	0.1321	0.1505
Ours	4	0.8167	0.1781	0.6890	**1.0660**	**0.1045**	**0.1112**

the student must choose $\mathbf{x}_t$ from $(\mathbf{x}_{<t}, c_{\leq t})$, while the teacher score is computed with access to the whole trajectory. CMD prevents this mismatch by restricting the teacher score for frame t to visual context and camera embeddings available up to that frame. With Prefix Scoring, this visual context is the cached generated prefix under which the target was produced.

4 Experiments

4.1 Training and Baselines

Training details and descriptions of the autoregressive and camera-controlled baselines are provided in suppl.

4.2 Quantitative Evaluation

Video Quality. For short-horizon I2V generation, we evaluate CMD with chunk sizes one and four on the VBench-I2V benchmark (Huang et al., 2025b), including five fixed seeds per image–prompt pair for every method. We report the aggregate Quality, I2V, and Total scores in table 2; full component breakdowns are provided in suppl. Our chunk-4 model achieves the best I2V (96.54) and Total (88.47) scores, improving over the strongest baseline by 1.18 and 0.84 percentage points, respectively while our chunk-1 model ranks second on all three aggregate metrics.

Table 2: Short-video VBench-I2V aggregates

Method	Chunk Size	Q	I2V	Total
CausVid	3	77.48	84.06	80.77
Self-Forcing	3	79.35	92.02	85.69
LongLive	3	78.64	91.22	84.93
Rolling Forcing	3	78.46	92.61	85.54
Context Forcing	3	80.16	90.66	85.41
LingBot-World	3	79.96	93.76	86.86
Causal Forcing	1	**81.56**	93.70	87.63
Causal Forcing++	1	79.35	95.36	87.35
Ours	1	80.47	96.44	88.46
Ours	4	80.40	**96.54**	**88.47**

Table 3: Long-video SANA-WM aggregates

Method	Chunk Size	Q	S	Total
LingBot-World	3	80.23	24.20	69.03
LongLive	3	80.04	23.86	68.80
Rolling Forcing	3	80.04	23.66	68.77
Context Forcing	3	80.72	23.38	69.25
Ours	1	**81.39**	**24.51**	**70.02**
Ours	4	80.84	24.50	69.57

Long-Video Quality. For long-horizon evaluation, we use the 80 images and original prompts from the Simple split of the SANA-WM benchmark (Zhu et al., 2026a). Each method is evaluated over 501 frames at 16 fps (∼30 seconds) and 480×832 resolution without camera input. We report aggregate VBench scores, following SANA-WM, in table 3; full component breakdowns are provided in suppl. Our chunk-1 model achieves the best score on every aggregate metric. Relative to the strongest external baseline in each column, it improves Quality by 0.67 percentage points over Context Forcing, Semantic by 0.31 points over LingBot-World, and Total by 0.77 points over Context Forcing. Our chunk-4 model ranks second on every aggregate metric.

Camera-Controlled Evaluation. For camera-controlled evaluation, we use the Simple and Hard splits of the SANA-WM benchmark (Zhu et al., 2026a). Each method generates eight 125-frame windows per scene, yielding 640 videos per split. Following SANA-WM (Zhu et al., 2026a), we report the arithmetic mean across the two splits for Quality, Semantic, and Total scores together with rotation, translation, and Camera Matrix Consistency (CamMC) errors in table 1; the split-specific results are provided in suppl. To compute the camera errors, we stitch the windows, recover camera poses with π^3 (Wang et al., 2025), and align the recovered trajectory to the target using Umeyama $\mathrm{Sim}(3)$ (Umeyama, 1991). Rotation error is the mean angular difference in degrees, translation error is the mean Euclidean distance, and CamMC is the mean Frobenius distance between the aligned 3×4 camera matrices; all three are lower-is-better. Our chunk-1 model achieves the best average Semantic and Total scores, while our chunk-4 model achieves the lowest average rotation, translation, and CamMC errors.

Table 4: SANA-WM camera ablations (Simple/Hard mean): Q, S, Total, Rot., Trans., CamMC.

Method	Camera	VBench Scores ↑			Camera Error ↓		
		Quality	Semantic	Total	Rot.	Trans.	CamMC
Bidir. teacher	PRoPE	0.6787	0.1138	0.5657	11.8308	0.5447	0.6806
Full CMD	PRoPE	0.7763	0.1714	0.6553	2.2470	0.4177	0.4313
Bidir. teacher	Ray map	0.7967	<u>0.1848</u>	0.6743	6.4290	0.2943	0.3531
Base CMD	Ray map	<u>0.8171</u>	0.1045	<u>0.6746</u>	**1.9436**	**0.1261**	**0.1442**
Full CMD	Ray map	**0.8222**	**0.2418**	**0.7062**	<u>1.9641</u>	<u>0.1321</u>	<u>0.1505</u>

4.3 Ablative experiments

Short video generation. We report aggregate short-video ablations in table 5, with full component breakdowns provided in suppl. We compare (i) *Bidir. teacher*: the bidirectional teacher jointly scores the generated clip and can attend to both past and future frames; and (ii) *Base CMD*: the causal teacher uses previous noisy DMD frames for context instead of a separately cached prefix. Adding Prefix Scoring (PS) to Base CMD gives (iii) + PS_0, which uses clean cached student prefixes. Full CMD combines PS with Prefix Corruption: (iv) *Full* CMD_{128} uses weaker prefix noise; (v) *Full* CMD_{512} uses stronger noise; and (vi) *Full* CMD_{256} is our default. Subscripts denote the scheduler timestep t_{prefix}, which determines ρ; this parameter does not apply to Base CMD. Moving from bidirectional scoring to Base CMD raises Total by 5.68 points and places Base CMD second on all aggregate metrics. Our default Full CMD further improves Total by 0.13 points.

Long video generation. We report aggregate long-video ablations in table 6, with full component breakdowns provided in suppl. We compare (i) *Bidir. teacher*: the score teacher processes each clip with bidirectional attention and no preceding rollout context; (ii) *Base CMD*: the teacher is causal within each clip, but receives no cross-clip prefix; and (iii) *Full CMD*: the causal teacher receives the same bounded cross-clip prefix as the student. Here, the prefix carries the student's bounded history across clip boundaries. Base CMD improves Total by 3.39 points over bidirectional scoring, while Full CMD improves Total by 4.04 points and Semantic by 0.58 points.

Camera control. We report chunk-1 camera-control ablations in table 4 as Simple/Hard means; split-specific results appear in suppl. We compare (i) *Bidir. teacher + PRoPE*: a bidirectional teacher with PRoPE ; (ii) *Full CMD + PRoPE*: a causal teacher with PRoPE; (iii) *Bidir. teacher + Ray map*: a bidirectional teacher with ray maps; (iv) *Base CMD + Ray map*: a causal teacher conditioned on noisy prior DMD frames rather than a cached prefix; and (v) *Full CMD + Ray map*: a causal teacher with ray maps and the cached student prefix. With PRoPE, our causal formulation improves every averaged metric over bidirectional scoring. With ray maps, Base CMD improves average Quality by 0.0204 over bidirectional scoring and reduces all three average camera errors. Adding the cached prefix improves average Total by a further 0.0316, while Base CMD retains slightly lower average camera errors.

Table 5: Short-video VBench-I2V ablations

Method	Q	I2V	Total
Bidir. teacher	74.48	90.81	82.65
Base CMD	<u>80.25</u>	<u>96.40</u>	<u>88.33</u>
+ PS_0	79.67	94.83	87.25
Full CMD_{128}	79.90	96.18	88.04
Full CMD_{512}	79.89	96.23	88.06
Full CMD_{256}	**80.47**	**96.44**	**88.46**

Table 6: Long-video SANA-WM ablations

Method	Q	S	Total
Bidir. teacher	76.49	<u>23.93</u>	65.98
Base CMD	<u>80.75</u>	23.86	<u>69.37</u>
Full CMD	**81.39**	**24.51**	**70.02**

4.4 Qualitative Analysis

Visual comparisons. Figure 5 compares representative rollouts across both evaluation settings. In the long-video example, our model preserves the first-person vehicle structure while progressing through distinct indoor spaces. In the short-video example, our model keeps the horse coherent while moving it forward along the rooftop towards the dome.

Automated pairwise preference. We compare our chunk-1 and chunk-4 models with six baselines on 100 VBench-I2V examples using a blind, forced-choice Gemini 3.1 Pro Preview judge. Paired videos share the same prompt and reference image, with judging prioritizing requested action and camera-motion fulfillment. In fig. 6, Chunk-1 and chunk-4 win 60-85% and 64-88% of comparisons, respectively. Full evaluation details are provided in suppl.

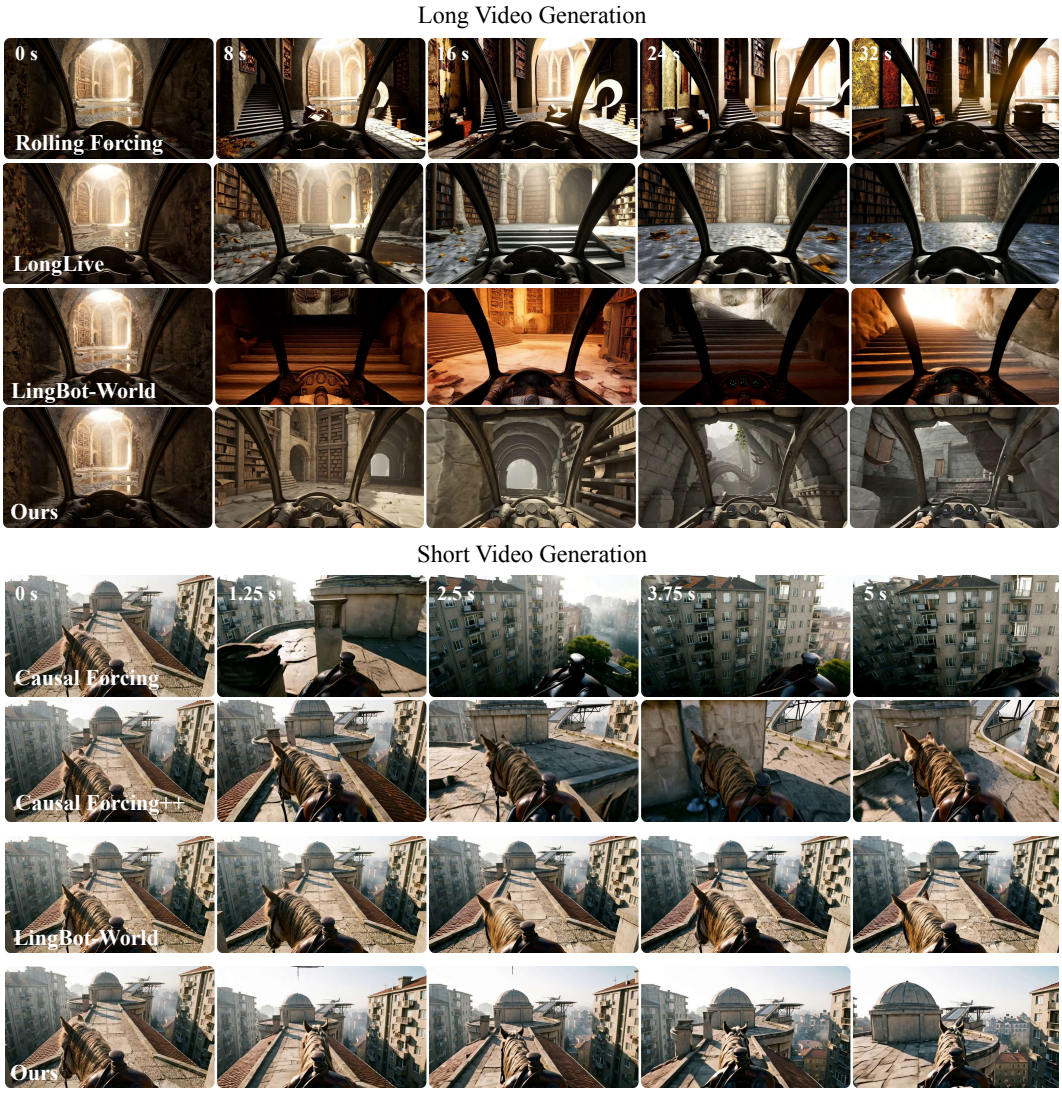


Figure 5: Qualitative comparisons for long-video (top) and short-video (bottom) generation.

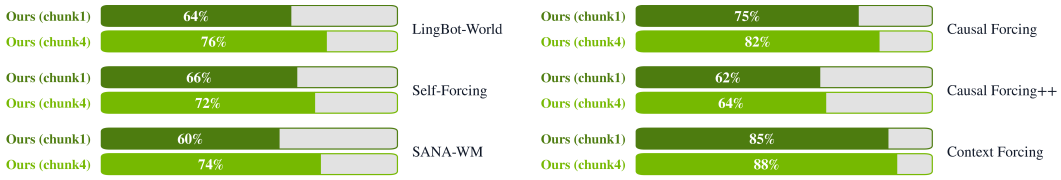


Figure 6: LLM preference study: Comparing chunk-1 and chunk-4 models.

5 Conclusion

We present Context-Matched Distillation (CMD), a causal DMD framework for few-step autoregressive video generation. CMD addresses the teacher–student information-set mismatch by ensuring that teacher scores cannot depend on future frames or controls. We use the same causal model family for teacher training, student initialization, and inference, avoiding a separate task-specific bidirectional teacher. Within CMD, Prefix Scoring evaluates each target using the student-generated history and controls available when that target was produced. Prefix Corruption makes this supervision robust to early student rollouts with severe artifacts. The same causal formulation supports long-video and camera-controlled generation with frame-relative conditioning. Across short and long rollouts, CMD improves generation quality over bidirectional scoring and substantially reduces camera-pose errors. Our ablations isolate the benefits of causal scoring, matched rollout context, and prefix corruption. Overall, CMD shows that matching supervision to the student's causal information state provides a practical foundation for efficient, long-horizon, and controllable video generation.

AI USE STATEMENT

Generative AI was used to generate a portion of the training data, for automated pairwise evaluation, and for language polishing. The authors reviewed all AI-assisted work and take responsibility for the final content.

REFERENCES

Hmrishav Bandyopadhyay, Rahim Entezari, Jim Scott, Reshinth Adithyan, Yi-Zhe Song, and Varun Jampani. Sd3.5-flash: Distribution-guided distillation of generative flows. In *ECCV*, 2026.

Shengqu Cai, Weili Nie, Chao Liu, Julius Berner, Lvmin Zhang, Nanye Ma, Hansheng Chen, Maneesh Agrawala, Leonidas Guibas, Gordon Wetzstein, et al. Mode seeking meets mean seeking for fast long video generation. *arXiv preprint arXiv:2602.24289*, 2026.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *NeurIPS*, 2024.

Jiaqi Feng, Justin Cui, Yuanhao Ban, and Cho-Jui Hsieh. One-forcing: Towards stable one-step autoregressive video generation. *arXiv preprint arXiv:2605.23458*, 2026.

Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing. *arXiv preprint arXiv:2411.16375*, 2024.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

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.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In *ICCV*, 2025.

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*, 2025a.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE TPAMI*, 2025b.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *NeurIPS*, 2025.

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.

Xiaofeng Mao, Shaohao Rui, Kaining Ying, Bo Zheng, Chuanhao Li, Mingmin Chi, and Kaipeng Zhang. Packforcing: Short video training suffices for long video sampling and long context inference. *arXiv preprint arXiv:2603.25730*, 2026.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE TPAMI*, 1991.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH*, 2024.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. In *NeurIPS*, 2025.

Youcan Xu, Jiaxin Shi, Zhen Wang, Wensong Song, Feifei Shao, Chen Liang, Jun Xiao, and Long Chen. Realcam: Real-time novel-view video generation with interactive camera control. *arXiv preprint arXiv:2605.06051*, 2026.

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.

Hidir Yesiltepe, Tuna Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. In *CVPR*, 2026.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024b.

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 *CVPR*, 2025.

Yifei Yu, Xiaoshan Wu, Xinting Hu, Tao Hu, Yangtian Sun, Xiaoyang Lyu, Bo Wang, Lin Ma, Yuewen Ma, Zhongrui Wang, et al. Videossm: Autoregressive long video generation with hybrid state-space memory. *arXiv preprint arXiv:2512.04519*, 2025.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *NeurIPS*, 2025a.

Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhengzhong Liu, and Hao Zhang. Fast video generation with sliding tile attention. *arXiv preprint arXiv:2502.04507*, 2025b.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rcm: A unified teacher-forcing and self-forcing open recipe for autoregressive diffusion distillation in streaming video generation and interactive world models. *arXiv preprint arXiv:2606.25473*, 2026.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. SANA-WM: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026a.

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*, 2026b.

Junhao Zhuang, Shiyi Zhang, Yuxuan Bian, Yaowei Li, Yawen Luo, Yijun Liu, Weiyang Jin, Songchun Zhang, Xianglong He, Xuying Zhang, et al. Self gradient forcing: Native long video extrapolation. *arXiv preprint arXiv:2607.20368*, 2026.