LONGTAKE: LEARNING TO SUSTAIN DYNAMICS IN LONG-HORIZON VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

World models, game simulators, and long-take video creation require coherent scene evolution and sustained dynamics over extended durations. Autoregressive (AR) video diffusion provides a natural framework for long-horizon generation, yet extended rollouts often become near-static or lose visual quality. We hypothesize that these failures reflect the limited guidance provided by short-video supervision on how ongoing scene dynamics develops over longer durations. This motivates us to introduce *LongTake*, a two-stage training pipeline built around Long-Horizon Teacher Forcing (TF) on curated real long videos. Long-Horizon TF trains the AR model to predict later frames conditioned on long ground-truth video prefixes, extending direct supervision beyond the short training horizon. This supervision is designed to help the model sustain dynamics and preserve visual quality during long-horizon generation. Our central finding is that this training stage strengthens direct initialization for distribution matching distillation (DMD) under student self-rollout, without the intermediate few-step distillation stage used in standard pipelines. Under the same five-second DMD training setup, our initialization yields substantially higher dynamic degree than short horizon TF initialization on 30-second rollouts at comparable aesthetic quality, and surpasses the evaluated baselines in both measures. Hybrid DMD further reuses this teacher to extend supervision to later frames of the self-rollout while retaining bidirectional joint supervision over the initial window. On long-horizon self-rollouts, *LongTake* lies on the Pareto front of dynamic degree and aesthetic quality, and Hybrid DMD attains the highest dynamic degree among evaluated methods at both 30s and 60s.

Project page : https://longtake-anonymous.github.io/

1 INTRODUCTION

Recent advances in video generation have enabled visually compelling synthesis (Yang et al., 2025; Kong et al., 2024; Wan et al., 2025), supporting applications in world modeling (Mao et al., 2026a; Sun et al., 2026; Hong et al., 2025; Chen et al., 2026b), game simulation (Tang et al., 2025; Qian et al., 2026), and interactive content creation (Shin et al., 2026; Huang et al., 2025b; Ki et al., 2026; Xiao et al., 2025; Zhang et al., 2026). In interactive simulation, generation needs to keep responding to successive actions (Yang et al., 2024) throughout multi-minute sessions (Valevski et al., 2025). Similarly, a requested long take (Xiong et al., 2024) of "A stylish woman strolls down a bustling Tokyo street." calls for her walk to continue coherently throughout the shot. A compelling opening is insufficient if activity fades into near-static frames or visual quality deteriorates over time.

Autoregressive (AR) video diffusion generates successive chunks conditioned on a key–value cache (KV-cache) of the video prefix (Yin et al., 2025; Teng et al., 2025; Chen et al., 2025). Although many AR models are trained on five-second clips (Huang et al., 2025a; Zhu et al., 2026), these applications require generation beyond the short training horizon (*e.g.*, minutes or even an hour). Recent methods stabilize such generation through context management (Yi et al., 2026; Yesiltepe et al., 2026; Liu et al., 2026; Kim et al., 2026; Mao et al., 2026b) and guidance from the video prefix (Song et al., 2025). However, generated scenes can still lose visual quality (Cui et al., 2026; Li et al., 2026a) or settle into near-static or repetitive states (Henschel et al., 2025; Lu et al., 2026; Dalva & Yanardag, 2026). Such stagnation is easy to miss, as consistency metrics can favor low-dynamic videos even when the requested dynamics are no longer sustained (Liao et al., 2024; Huang et al., 2024).

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.

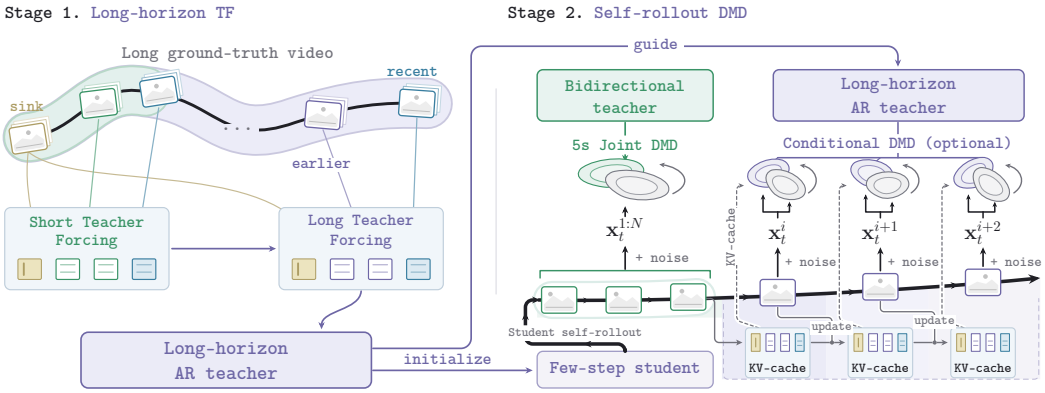

Figure 1: **LongTake learns long-horizon generation through a two-stage training pipeline.** (Left) *Stage 1:* Long-Horizon TF learns from real long videos and provides a strong direct initialization for self-rollout DMD. (Right) *Stage 2:* Self-rollout DMD uses this initialization. Hybrid DMD further adds conditional supervision of later frames from the AR teacher, while retaining joint supervision from the frozen bidirectional teacher over the initial window of the student rollout.

These failures motivate learning long-horizon generation from real long videos, which show how a scene keeps evolving while remaining visually coherent over tens of seconds (Xiong et al., 2024). Prior work shows that extending the temporal span of training sequences can improve the modeling of temporal dynamics, since sequences shorter than an effect provide little training signal for it (Brooks et al., 2022). Training on short video clips thus provides no direct supervision for effects that unfold over tens of seconds and leaves the model to generalize beyond its training horizon. Pairing long video prefixes with the frames that follow them instead extends supervision to later stages of the same evolving scene, where long-horizon generation takes place. This leads to the following question.

Can supervision from real long videos help sustain coherent dynamics while preserving visual quality as the scene evolves during long-horizon autoregressive rollout?

We introduce *LongTake*, a two-stage training pipeline built around Long-Horizon Teacher Forcing (TF) on curated real long videos as described in Figure 1. Here, Long-Horizon TF constructs the conditioning KV-cache from long ground-truth video prefixes and directly supervises later frames beyond the short training horizon. Our central finding is that the resulting model serves as a strong direct initialization for self-rollout distribution matching distillation (DMD) (Zhuang et al., 2026), supporting effective two-stage training without a separate few-step initialization stage (Zhu et al., 2026; Zhao et al., 2026). Under the same 5s DMD setting (Zhuang et al., 2026), initializing from the Long-Horizon TF model instead of a standard short-clip TF model raises the VBench dynamic degree (Huang et al., 2024) of 30s rollouts while preserving comparable aesthetic quality and also surpasses the three-stage pipeline in both measures. Beyond initialization, extending DMD to later frames relies on teacher scores under longer video prefixes, where a 5s DMD teacher may provide limited guidance. Hybrid DMD therefore reuses the Long-Horizon TF model to provide conditional scores for later frames of the student self-rollout. On 30s and 60s rollouts, both LongTake variants lie on the Pareto front of dynamic degree against aesthetic quality and against non-motion VBench quality, as the only methods on either front above 90 in dynamic degree.

Our contributions are summarized below.

- We curate real long videos and introduce Long-Horizon TF to supervise later frames from long video prefixes beyond the short training horizon.
- We show that Long-Horizon TF on curated real long videos strengthens initialization for the same 5s self-rollout DMD, yielding stronger long-horizon dynamics at comparable visual quality without the separate few-step initialization stage of the standard pipeline.
- We introduce Hybrid DMD, which reuses the Long-Horizon TF model as a teacher for later frames of the student self-rollout and attains the highest dynamic degree.
- On 30s and 60s rollouts, *LongTake* lies on the Pareto front of dynamic degree and aesthetic quality with dynamic degree above 90, whereas baselines with higher aesthetic quality stay at or below 58.

2 Related Work

Long-video supervision and AR initialization. Real long videos provide direct supervision for scene evolution beyond short clips. LVD-2M (Xiong et al., 2024) and Presto (Yan et al., 2025) emphasize dynamic long-take videos, while Long Context Tuning (Guo et al., 2025) learns dependencies across extended scenes. For AR initialization, Causal Forcing (Zhu et al., 2026) and Causal Forcing++ (Zhao et al., 2026) use separate causal ODE and consistency distillation stages before self-rollout DMD. SGF (Zhuang et al., 2026) also evaluates direct initialization from a TF-trained model. Building on this approach, LongTake studies how supervision from curated real long videos strengthens direct initialization. To this end, Long-Horizon TF supervises later frames under KV-cache states constructed from long video prefixes. Under the same 5s joint DMD, this initialization yields stronger long-horizon dynamics while preserving comparable visual quality.

Long-horizon self-rollout distillation Self Forcing (Huang et al., 2025a) trains under student self-rollout to reduce the *training–inference gap*. Self-Forcing++ (Cui et al., 2026) applies DMD to short windows from extended student rollouts, while Cai et al. (2026) combine long-video supervision with sliding-window distribution matching. Context Forcing (Chen et al., 2026a) also trains a long-context teacher on real videos and jointly matches a target window under the preceding student self-rollout. Context-Matched Distillation (Bandyopadhyay et al., 2026) initializes the student from its causal teacher and scores each chunk under its own student-generated prefix. Hybrid DMD scores each next chunk in the same way but reuses the Long-Horizon AR teacher and retains bidirectional joint supervision over the initial window. Further comparisons appear in Appendix A.

3 Preliminaries

In this section, we review the three-stage training pipeline commonly used for few-step AR video diffusion (Zhu et al., 2026; Zhao et al., 2026; Zheng et al., 2026). As summarized in Figure 2, AR diffusion training is followed by a separate few-step initialization stage and self-rollout DMD. We use this pipeline as a reference to examine how Long-Horizon TF strengthens direct initialization for DMD.

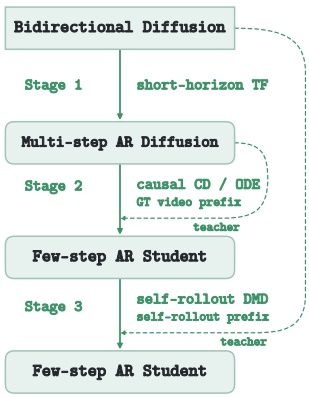

Figure 2: Three-stage pipeline for few-step AR video diffusion (Zheng et al., 2026).

Stage 1: AR diffusion training AR video diffusion generates a video as a sequence of chunks[1], with the AR factorization

$$p(\mathbf{x}) = \prod_{i=1}^{N} p^{\mathtt{AR}}(\mathbf{x}^i|\mathbf{x}^{<i}), \tag{1}$$

where $\mathbf{x}^i$ denotes a chunk of consecutive latent frames. A common starting point for AR diffusion distillation is to fine-tune a pretrained bidirectional diffusion model with teacher forcing (TF) (Jin et al., 2025). Each noisy chunk $\mathbf{x}_t^i$ at a sampled noise level t is conditioned on the keys and values computed by the AR model from the preceding clean ground-truth video prefix $\mathbf{x}_{\mathtt{gt}}^{<i}$. This gives the standard AR flow-matching objective for short-horizon TF

$$\phi^\star = \arg\min_{\phi} \mathbb{E}_{i\in[1,N],t_i,\varepsilon_i,\mathbf{x}_{\mathtt{gt}}\sim\mathcal{D}} \left[w(t_i) \left\| v_\phi^{\mathtt{AR}}(\mathbf{x}_{t_i}^i, \mathbf{x}_{\mathtt{gt}}^{<i}, \varnothing, t_i) - (\varepsilon_i - \mathbf{x}_{\mathtt{gt}}^i) \right\|^2 \right], \tag{2}$$

where $\mathbf{x}_{t_i}^i = (1-t_i)\mathbf{x}_{\mathtt{gt}}^i + t_i\varepsilon_i$ and $\varepsilon_i \sim \mathcal{N}(0, I)$. Standard TF in (2) is applied to a short ground-truth clip $\mathbf{x}_{\mathtt{gt}} \sim \mathcal{D}_{\mathtt{short}}$ containing at most N frames (*e.g.*, 21 latent frames). The model receives the clean video prefix $\mathbf{x}_{\mathtt{gt}}^{<i}$ and an empty initial KV-cache $\varnothing$. Using causal masking, the AR model evaluates losses for all target chunks in parallel in one forward pass.

Stage 2: Few-step initialization The multi-step AR diffusion model $v_{\phi^\star}^{\mathtt{AR}}$ from (2) can be further converted into a few-step generator $v_{\theta^\star}^{\mathtt{AR}}$ before self-rollout training. Existing pipelines perform this initialization using ODE distillation (Zhu et al., 2026), consistency distillation (Zhao et al., 2026; Zheng et al., 2026), where the AR teacher and student use the same ground-truth video prefix from $\mathbf{x}_{\mathtt{gt}} \sim \mathcal{D}_{\mathtt{short}}$. This stage provides a few-step initialization for the subsequent self-rollout distribution-matching stage using student-generated videos.

[1]We use chunk-level notation throughout the paper, with frame-wise AR as the single-frame chunk case.

Stage 3: Distribution matching distillation Few-step initialization uses ground-truth video prefixes $\mathbf{x}_{\mathtt{gt}}^{<i}$. At inference, the conditioning KV-cache is instead updated from student-generated chunks. Self-rollout training (Huang et al., 2025a) reduces this gap by generating according to the factorization

$$q^{\theta}(\mathbf{x}_{\mathtt{self}}) = \prod_{i=1}^{N} q^{\theta}\left(\mathbf{x}_{\mathtt{self}}^{i} | \mathcal{C}_{\mathtt{cache}}(\mathbf{x}_{\mathtt{self}}^{<i})\right), \tag{3}$$

where $\mathcal{C}_{\mathtt{cache}}$ maps a video prefix to a bounded per-layer KV-cache that retains keys and values from preceding chunks. This KV-cache is updated after each generated chunk according to the cache policy. Distribution matching distillation (DMD) (Yin et al., 2024; Huang et al., 2025a) uses a frozen bidirectional teacher $p_t^{\star}$ to supervise the joint distribution q_t^{θ} of the student rollout through:

$$\mathcal{J}_{\mathtt{DMD}}(\theta) = \mathbb{E}_{t, \mathbf{x}_{\mathtt{self}} \sim q^{\theta}}\left[w(t) D_{\mathtt{KL}}\left(q_t^{\theta}(\mathbf{x}_t) | p_t^{\star}(\mathbf{x}_t)\right)\right]. \tag{4}$$

Hence, DMD in (4) trains the generator on self-rollouts $\mathbf{x}_{\mathtt{self}}$, exposing it to KV-cache constructed from its own generated chunks and reducing the *training–inference gap* (Huang et al., 2025a).

4 Training for Long-Horizon Video Generation

In this section, we first introduce Long-Horizon TF to learn long-horizon generation from real long videos. Our central finding is that the resulting initialization substantially improves long-horizon generation under the same short-horizon self-rollout DMD procedure, enabling effective two-stage training. Hybrid DMD further uses this AR teacher to supervise later frames of the student rollout.

4.1 Long-Horizon Supervision for Autoregressive Generation

Let N denote the short-video training horizon and $M \gg N$ the target rollout horizon. The standard pipeline in Figure 2 trains the AR diffusion model and performs self-rollout DMD over N (*e.g.*, 21 latent frames). During AR training, the ground-truth video prefix $\mathbf{x}_{\mathtt{gt}}^{<i}$ is encoded into the conditioning KV-cache, and learning from short clips directly supervises $p_{\phi}^{\mathtt{AR}}(\mathbf{x}^i \mid \mathbf{x}_{\mathtt{gt}}^{<i})$ only for $i < N$. Long-horizon generation, however, requires the model to continue from longer video prefixes at positions $i \geq N$. Because short-clip training provides no direct supervision for prediction at these later positions, the model relies on generalization to continue beyond the short-video training horizon through the AR self-rollout.

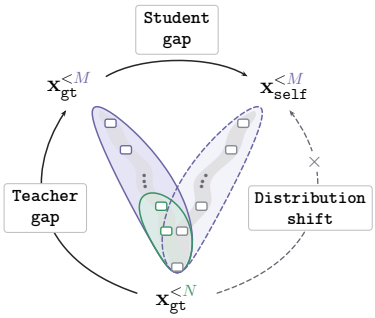

Figure 3: **Teacher and student gaps in KV-cache states.**

Teacher gap The *teacher gap* concerns the shift from KV-cache states constructed from ground-truth videos within N to those encountered toward M. Real long videos provide the missing supervision by pairing later conditioning KV-cache states with corresponding ground-truth target frames. Long-Horizon TF trains the AR teacher on these pairs to generate later frames, as detailed in Section 4.2.

Student gap Long-Horizon TF expands teacher supervision to KV-cache states constructed from ground-truth videos throughout $M \gg N$. At inference, however, the conditioning KV-cache is constructed from student self-rollout $\mathbf{x}_{\mathtt{self}}^{<i}$. These states can differ from those constructed from the ground-truth video prefixes $\mathbf{x}_{\mathtt{gt}}^{<i}$ used in TF. We refer to this familiar *training–inference gap* (Huang et al., 2025a) as the *student gap*. Self-rollout DMD addresses this mismatch by training under student-generated video prefixes. Figure 3 summarizes the teacher gap and student gap, which concern the training horizon and the transition to student self-rollout, respectively.

4.2 Long-Horizon Teacher Forcing

We extend teacher forcing to prediction positions beyond N using real long videos $\mathbf{x}_{\mathtt{gt}}$ from the curated dataset $\mathcal{D}_{\mathtt{Long}}$. Long-Horizon TF constructs the conditioning KV-cache by applying the chosen cache update rule to longer ground-truth video prefixes. Supervision on the later target frames teaches the AR model to predict them from the temporal context preserved as the same scene evolves over extended durations, which targets the *teacher gap*.

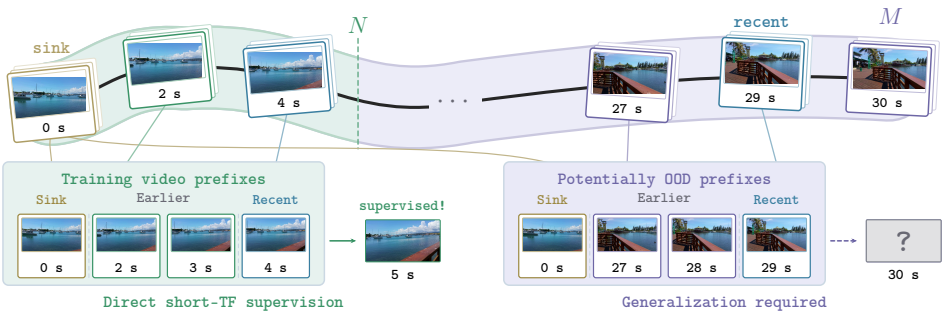

Figure 4: **Teacher gap with a bounded KV-cache.** With sink retention, the temporal separation between sink frames and recent frames increases as the video prefix grows. Short-horizon TF directly supervises prediction only within N, while generating later frames toward M relies on generalization.

Long-Horizon TF Given a ground-truth video $\mathbf{x}_{\mathtt{gt}} \sim \mathcal{D}_{\mathtt{Long}}$, we sample a valid split s and apply TF to a target window of length N beginning at s. The conditioning KV-cache at the split is $\mathtt{c}^s_{\mathtt{gt}} := \mathcal{C}_{\mathtt{cache}}(\mathbf{x}^{<s}_{\mathtt{gt}})$, where the chosen cache update rule determines which information from the video prefix is preserved. With sink retention (Yang et al., 2026), the initial frames remain in the conditioning KV-cache while the recent frames are updated as the video prefix grows. Sampling longer video prefixes therefore provides supervision over larger temporal separations between the retained early and recent frames, covering how later frames relate to this context. For each target chunk at $i \in [s, s+N)$, the model uses this cache and preceding ground-truth frames $\mathbf{x}^{s<i}_{\mathtt{gt}}$ within the target window, giving the flow-matching objective defined as follows:

$$\phi^\star = \arg\min_{\phi} \mathbb{E}_{s,i\in[s,s+N),t_i,\varepsilon_i,\mathbf{x}_{\mathtt{gt}}\sim\mathcal{D}_{\mathtt{Long}}} \left[w(t_i) \left\| v^{\mathtt{AR}}_{\phi}\left(\mathbf{x}^i_{t_i}, \mathbf{x}^{s<i}_{\mathtt{gt}}, \mathtt{c}^s_{\mathtt{gt}}, t_i\right) - \left(\varepsilon_i - \mathbf{x}^i_{\mathtt{gt}}\right) \right\|^2 \right]. \quad (5)$$

When $s = 0$, the conditioning KV-cache is empty, *i.e.*, $\mathtt{c}^s_{\mathtt{gt}} = \varnothing$, recovering standard TF in (2).

Parallel cache computation The entire conditioning KV-cache can be constructed in parallel for both sink retention with FIFO eviction (Yang et al., 2026). Specifically, we compute prefix Q/K/Vs in parallel with causal masking and select retained entries according to their temporal positions. For the EMA extension (Lu et al., 2026), a parallel prefix scan (Blelloch, 1990) incorporates the accumulated contribution of evicted entries into the sink. Both variants therefore produce $\mathtt{c}^s_{\mathtt{gt}}$ for each layer within the same parallel cache-construction pass. Then, next pass computes TF losses over the fixed target window as in (2). The clean stream supplies preceding ground-truth frames, while the noisy stream predicts target velocities. For $s > 0$, Long-Horizon TF thus separates cache construction from target prediction into *two forward passes*, as summarized in Algorithm 1.

Algorithm 1 Long-Horizon TF

Note: Details of parallel computation are in Appendix B.1.

```
# v - autoregressive flow model
# C - bounded KV-cache
# N - short TF horizon
while True:
    x = sample(D_Long)
    M = len(x)
    s = randint(0, M - N)

    # Pass 1: Parallel cache prefill
    prefix = x[:s]
    c = parallel_prefill(prefix)

    # Pass 2: Teacher-forcing
    x_tar = x[s:s+N]
    t = sample_timestep()
    eps = randn_like(x_tar)
    x_t = (1 - t) * x_tar + t * eps
    v_tar = eps - x_tar
    v_hat = v(x_t, x_tar, c, t)

    loss = mse_loss(v_hat, v_tar)
    update(v, loss)
return v
```

Effect of Long-Horizon TF initialization Following SGF (Zhuang et al., 2026), we initialize self-rollout DMD directly from a TF-trained model while retaining the same DMD procedure and its short horizon N. In Figure 5, compared with short-horizon TF initialization, causal CD initialization produces stronger dynamics but lower aesthetic quality. Long-Horizon TF initialization improves dynamics while retaining comparable aesthetic quality to short-horizon TF initialization.

Finding. *Long-Horizon TF on curated real long videos strengthens direct initialization for self-rollout DMD. Under the same DMD procedure, the resulting student achieves stronger long-horizon dynamics with comparable visual quality, without a separate few-step initialization.*



Figure 5: **Long-Horizon TF enables effective two-stage training.** All models use 5s self-rollout DMD and are evaluated on 30s rollouts. Long TF initialization improves long-horizon dynamics over Short TF while maintaining comparable visual quality, showing that stronger TF initialization supports effective long-horizon generation without a separate CD stage.

4.3 Extending Supervision with Hybrid DMD

We next extend the benefit of Long-Horizon TF by reusing the resulting AR teacher to supervise later frames under student self-rollout. Hybrid DMD adds this conditional supervision while retaining joint supervision over the initial window from the frozen bidirectional teacher.

Extending supervision Self-rollout DMD (Huang et al., 2025a) trains the student using its own generated frames, but supervision over the initial window of length N provides no direct training signal for later frames. Long-Horizon TF in (5) trains the AR teacher $p^{\mathtt{AR}}_{\phi^\star}$ to generate frames beyond N. For each target chunk beyond the initial window, we construct its conditioning KV-cache $\mathtt{c}^i_{\mathtt{self}} := \mathcal{C}_{\mathtt{cache}}(\mathbf{x}^{<i}_{\mathtt{self}})$ from the student-generated video prefix. The teacher then provides a conditional target $p^{\mathtt{AR}}_{\phi^\star}(\cdot \mid \mathtt{c}^i_{\mathtt{self}})$ for that chunk. Whereas Context Forcing (Chen et al., 2026a) jointly scores a target window, our conditional DMD scores each chunk using the KV-cache updated from preceding student-generated frames. Appendix A.1 compares the two scoring objectives and their conditioning video prefixes at each target chunk.

Algorithm 2 Hybrid DMD

Note: Details of Hybrid DMD are in Appendix B.2.

```
# G - few-step AR generator
# C - bounded KV-cache policy
# N - initial joint-DMD horizon
while True:
    # Student self-rollout
    x, C = self_rollout(G)

    # Hybrid DMD
    l_jnt = joint_DMD(x[:N],∅,teacher=p^Bi)

    l_cnd = 0
    for i in range(N, len(x)):
        l_cnd += cond_DMD(x[i],C[i],
                teacher=p^AR)
    l_cnd /= len(x[N:])

    loss = l_jnt + λ * l_cnd
    update(G, loss)
return G
```

Hybrid DMD We retain the pretrained bidirectional teacher $p^{\mathtt{Bi}}$ for joint supervision over the initial window of length N and use the Long-Horizon AR teacher to supervise later frames. Together, they define the hybrid teacher distribution for student self-rollout over the horizon M as

$$p^\star(\mathbf{x}^{0:M}_{\mathtt{self}}) = p^{\mathtt{Bi}}(\mathbf{x}^{0:N}_{\mathtt{self}}) \prod_{i=N}^{M-1} p^{\mathtt{AR}}_{\phi^\star}\left(\mathbf{x}^i_{\mathtt{self}} \mid \mathtt{c}^i_{\mathtt{self}}\right). \tag{6}$$

We apply joint DMD over the initial window and conditional DMD beyond that window, with $\lambda > 0$ controlling the contribution of conditional supervision relative to joint supervision. Combining these terms gives the Hybrid DMD objective for joint and conditional distribution matching over M as

$$\begin{aligned}\mathcal{J}_{\mathtt{Hybrid-DMD}}(\theta) = {} & D_{\mathrm{KL}}\left(q^\theta(\mathbf{x}^{0:N}_{\mathtt{self}}) \,\middle\|\, p^{\mathtt{Bi}}(\mathbf{x}^{0:N}_{\mathtt{self}})\right) \\ & + \lambda \mathbb{E}_{i\in[N,M)} \mathbb{E}_{\mathbf{x}^{<i}_{\mathtt{self}} \sim q^\theta}\left[D_{\mathrm{KL}}\left(q^\theta(\mathbf{x}^i_{\mathtt{self}} \mid \mathtt{c}^i_{\mathtt{self}}) \,\middle\|\, p^{\mathtt{AR}}_{\phi^\star}(\mathbf{x}^i_{\mathtt{self}} \mid \mathtt{c}^i_{\mathtt{self}})\right)\right].\end{aligned} \tag{7}$$

Here, the expectation averages conditional matching over student-generated video prefixes $\mathbf{x}^{<i}_{\mathtt{self}}$ and the teacher and student construct their conditioning KV-cache from the same prefix while retaining their respective representations. Both teachers remain frozen throughout the joint and conditional supervision of student-generated frames, as summarized in Algorithm 2.

4.4 PROPOSED TRAINING PIPELINE

We combine Long-Horizon TF and self-rollout DMD into the two-stage training pipeline shown in Figure 6. The first stage trains an AR diffusion model on the real long-video dataset $\mathcal{D}_{\text{Long}}$ using Long-Horizon TF in (5). The resulting weights directly initialize the student for the second stage of self-rollout DMD, without a separate few-step initialization stage. We begin with short-horizon joint DMD, where the frozen bidirectional teacher supervises the joint distribution of student self-rollouts of length N. This completes the standard two-stage *LongTake* pipeline. As an extension within the second stage, Hybrid DMD continues training on longer student rollouts. The bidirectional teacher retains joint supervision over the initial window of length N, while the TF-trained AR model serves as the frozen Long-Horizon AR teacher for conditional supervision of later frames. The AR teacher conditions on the preceding student-generated video prefix, and both teachers remain frozen as the student is updated. Appendix B.2 details the student updates and the gradient replay procedure used to compute them.

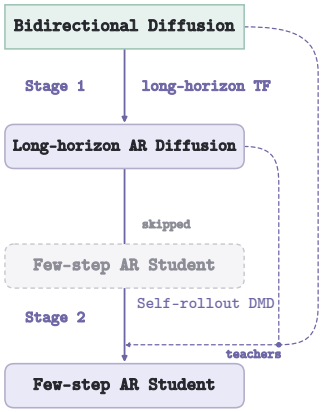

Figure 6: **Two-stage training**

5 EXPERIMENTS

Implementation details We use `Wan2.1-T2V-1.3B` (Wan et al., 2025) to generate videos at 832×480 and 16 FPS, with three latent frames per AR chunk and four denoising steps. Long-Horizon TF uses 43K video–text pairs, including OpenVidHD clips (Nan et al., 2025), and 3,000 updates to supervise generation up to 30s. The resulting model directly initializes self-rollout DMD using SGF (Zhuang et al., 2026) with extended VidProM prompts (Wang & Yang, 2024). LongTake uses 1,200 joint DMD iterations on 21-latent-frame rollouts. The Hybrid DMD variant uses 800 joint DMD iterations followed by 400 Hybrid DMD iterations on 42-latent-frame rollouts with $\lambda = 0.2$. Both variants use `Wan2.1-T2V-14B` as the bidirectional teacher and a fixed 12-latent-frame student attention window. Each variant uses its final EMA checkpoint for both evaluation durations. Appendix C provides checkpoint sources and complete training configurations for both stages.

Baselines and evaluation We compare publicly released `1.3B` checkpoints of Self Forcing (Huang et al., 2025a), Causal Forcing (Zhu et al., 2026), LongLive (Yang et al., 2026), Reward Forcing (Lu et al., 2026), MemRoPE (Kim et al., 2026), Rolling Forcing (Liu et al., 2026), SGF (Zhuang et al., 2026), and Context Forcing (Chen et al., 2026a) in Table 1. We evaluate 30s and 60s generation on 128 extended MovieGen prompts (Polyak et al., 2024) and 251 VBench-Long prompts (Huang et al., 2025c), respectively. Quality uses the standard VBench (Huang et al., 2024) normalization and weights across seven video-quality dimensions. We separately report Δ**IQ** (Liu et al., 2026) and **VLM** exposure (Kim et al., 2026) to assess long-horizon visual stability.

Qualitative results Figure 7 compares baselines with LongTake using Hybrid DMD. In Tokyo, the three methods with the lowest dynamic degree at 30s keep the woman in nearly the same pose and position from 8s to 30s, so the requested walk stalls even where frames stay sharp, whereas LongTake keeps her walking as the scene changes. At 60s, beyond the 30s supervision horizon, LongTake preserves the dog's appearance, while Causal Forcing develops color artifacts, MemRoPE duplicates the subject, and Rolling Forcing drifts into close-ups despite high dynamic degree.

Quantitative results As shown in Figure 8, the baselines trade motion for quality. Every baseline with higher aesthetic quality than LongTake reaches a dynamic degree of at most 58, while the only one above 90, Rolling Forcing, trails LongTake by 3.9 aesthetic score. LongTake lies on the Pareto front at both durations (see Appendix D.2) and, relative to other methods on the front, gains 33-42 scores of dynamic degree for under two aesthetic scores. Reward Forcing, the strongest baseline in Quality at 60s, matches LongTake in aesthetic quality with 14.5 fewer score of dynamic degree. In Table 1, LongTake with only 5s joint DMD also exceeds every baseline in Quality at both durations. Hybrid DMD achieves the highest dynamic degree at both durations and the highest Quality at 30s, while joint DMD retains higher Quality at 60s, so the preferred variant depends on the target duration.



Figure 7: **Long-horizon generation with sustained dynamics.** Qualitative comparisons over 30s (top) and 60s (bottom). Baseline rollouts exhibit appearance degradation or limited scene progression. LongTake with Hybrid DMD sustains scene evolution with coherent appearance in later frames.

Long-horizon visual stability We assess visual stability using Δ**IQ** (Liu et al., 2026) and **VLM** (Kim et al., 2026). They measure the absolute change in imaging quality between the first and last five seconds and exposure stability throughout the video, respectively. We interpret Δ**IQ** alongside imaging quality because stable endpoint scores alone do not establish high visual quality. LongTake matches the highest **VLM** score at 30s and achieves comparable Δ**IQ** to Context Forcing at both durations while producing substantially stronger dynamics. Both LongTake variants obtain lower Δ**IQ** than Rolling Forcing at 30s and 60s, although Hybrid DMD slightly increases endpoint quality changes relative to joint DMD alone. It also exceeds Rolling Forcing in **VLM** while producing stronger dynamics. These results support sustained dynamics with competitive visual stability beyond the teacher supervision horizon.

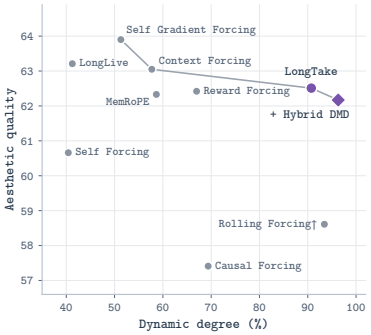

Figure 8: **Motion-quality trade-off at 30s.** Gray line: Pareto front.

Table 1: **Evaluation on 30s and 60s AR rollouts.** **Quality** aggregates seven video-quality dimensions with standard VBench normalization and weights. **ΔIQ** and **VLM** assess visual stability. Best and second-best results are highlighted for each duration. LongTake uses Long-Horizon TF followed by 5s joint DMD. Full results appear in Table 5. † checkpoint from (Zhu et al., 2026).

	30s						60s					
Method	**Subj.↑**	**Dyn.↑**	**Aesth.↑**	**Quality↑**	**ΔIQ↓**	**VLM↑**	**Subj.↑**	**Dyn.↑**	**Aesth.↑**	**Quality↑**	**ΔIQ↓**	**VLM↑**
Short-video supervision												
Self Forcing (Huang et al., 2025a)	97.43	40.42	60.66	82.34	5.82	3.13	97.27	44.21	56.58	81.89	6.65	2.69
Causal Forcing (Zhu et al., 2026)	96.10	69.38	57.41	82.19	7.34	2.45	95.96	59.86	51.83	80.93	12.97	2.07
LongLive (Yang et al., 2026)	98.03	41.25	63.21	83.25	2.96	3.90	98.40	37.04	62.23	82.98	2.69	3.83
Reward Forcing (Lu et al., 2026)	97.26	66.98	62.42	84.36	2.58	3.79	97.29	80.23	62.08	85.54	2.45	3.75
MemRoPE (Kim et al., 2026)	97.28	58.65	62.33	83.95	2.95	3.83	97.48	56.20	59.34	83.92	3.69	3.91
Rolling Forcing† (Liu et al., 2026)	96.04	93.44	58.61	84.67	5.58	3.77	95.81	95.69	56.75	84.78	6.69	3.80
Self Gradient Forcing (Zhuang et al., 2026)	98.19	51.30	63.90	84.25	2.75	3.94	98.42	53.06	64.09	84.15	2.70	3.87
Long-video supervision												
Context Forcing (Chen et al., 2026a)	97.72	57.71	63.05	84.16	3.31	3.55	98.17	52.69	62.10	83.85	3.65	3.26
LongTake (Ours)	96.83	90.73	62.51	85.79	3.13	3.94	96.78	94.72	62.35	85.96	3.55	3.87
+ Hybrid DMD ($\lambda = 0.2$)	96.32	96.30	62.17	85.97	3.22	3.92	96.43	97.36	61.48	85.71	3.60	3.86

Remark. *Videos with low dynamic degree show little subject or camera motion over time, as in Figure 7. Such videos no longer depict the action described in the prompt and extend duration without adding content, which undermines long-horizon generation. Yet they can still score well, since each frame stays sharp and consistency metrics favor static videos (Liao et al., 2024; Huang et al., 2024), so long videos need high dynamic degree as much as visual quality.*

Ablation Study We compare generation from shared video prefixes before DMD in Appendix D.1. Compared to standard TF, Long-Horizon TF improves quality at each prefix length, while boundary metrics suggest better appearance preservation at the immediate transition.

Table 2: **Hybrid DMD with shared initialization.**

λ	Subj.↑	Dyn.↑	Aesth.↑	Quality↑
0.0	96.83	90.73	62.51	85.79
0.1	96.78	90.00	62.64	85.83
0.2	96.32	96.30	62.17	85.97
0.3	96.38	95.57	61.49	85.85
0.4	96.39	95.42	61.91	85.61

Table 2 evaluates supervision of later frames on 30s generation. All settings use the same Long-Horizon TF initialization and 1,200 distillation iterations. Joint DMD retains 5s supervision throughout, while Hybrid DMD extends the rollout after 800 iterations. With $\lambda = 0.2$, Hybrid DMD achieves the highest dynamic degree and Quality score among the tested settings, with modest reductions in subject consistency and aesthetic quality. The 60s results in Table 1 show stronger dynamics with a small decrease in Quality relative to joint DMD, making Hybrid DMD an extension with a motion–quality tradeoff. Appendix D.3 illustrate the additional scene progression with coherent subject appearance.

6 Conclusion

We introduced *LongTake*, a two-stage training pipeline for long-horizon AR video generation. Long-Horizon TF learns from curated real long videos by predicting later frames conditioned on long ground-truth video prefixes. The resulting model directly initializes self-rollout DMD without a separate few-step initialization stage. Under the standard DMD, this initialization yields higher dynamic degree than short TF initialization at comparable aesthetic quality. Hybrid DMD further reuses the model as a teacher for later frames of the student self-rollout, achieving the highest dynamic degree among evaluated methods. Both variants lie on the Pareto front of dynamic degree and aesthetic quality on 30s and 60s rollouts. These results demonstrate the value of long-video supervision for learning to sustain dynamics over long-horizon.

Limitations The effectiveness of Hybrid DMD depends on the predictive quality of the Long-Horizon AR teacher used to supervise later frames. Future work could incorporate error-recycling fine-tuning, as in SVI (Li et al., 2026b), to improve teacher robustness to errors accumulated in the video prefix during student self-rollout. Evaluation is limited to a 1.3B backbone and rollouts up to 60s, leaving performance with larger backbones and longer rollouts untested.

AI USE STATEMENT

The research ideas and proposed methods originated with the authors. We used generative AI tools to help refine author-developed hypotheses, assist with implementing the proposed methods and experimental procedures, support dataset cleaning and reformatting, and assist with interpreting experimental results. Additionally, we used these tools for manuscript drafting and revision, as well as literature search, discovery, and summarization. We also used Gemini 3.1 Pro Preview to evaluate exposure stability in generated videos, following the protocol described in Appendix C.4. The authors reviewed and revised the AI-assisted material and take full responsibility for the final content of this work, including its methodology, implementation, claims, and reported results.

REPRODUCIBILITY STATEMENT

Appendix B.2 documents student rollout, gradient replay, and Hybrid DMD updates. Appendix C.1 specifies initialization, optimization, cache configurations, and sampling. Appendix C.2 describes the training corpus and curation, Appendix C.3 lists baseline checkpoints and generation settings, and Appendix C.4 defines the evaluation prompts, metrics, and score aggregation.

REFERENCES

Hmrishav Bandyopadhyay, Xuanchi Ren, Zijian Huang, Jay Zhangjie Wu, Tianshi Cao, Ruilong Li, Bryan Chu, Sanja Fidler, Yi-Zhe Song, and Zian Wang. Context-matched distillation: Teacher causality for autoregressive video distillation. *arXiv preprint arXiv:2608.13391*, 2026.

Guy E Blelloch. Prefix sums and their applications. 1990.

Tim Brooks, Janne Hellsten, Miika Aittala, Ting chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei A Efros, and Tero Karras. Generating long videos of dynamic scenes. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=VnAwNNJiwDb.

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 *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=kVBNDXHVeo.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=yDo1ynArjj.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

Shuo Chen, Cong Wei, Sun Sun, Tiancheng SHEN, Ping Nie, Kai Zou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=lv86f7Qr9V.

Zhifei Chen, Luozhou Wang, Guibao Shen, Dongyu Yan, Shuai Yang, Tianshuo Xu, Yihua Du, Wei Wang, Tianyi Gui, Lianghua Huang, et al. Reworld: An interactive world model with long-horizon memory. *arXiv preprint arXiv:2608.23565*, 2026b.

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. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DzvPiqh23f.

Yusuf Dalva and Pinar Yanardag. Adastate: Self-evolving anchors for streaming video generation. *arXiv preprint arXiv:2605.30349*, 2026.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3674–3683. IEEE, 2020.

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*, pp. 17281–17291, 2025.

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. In *European Conference on Computer Vision*, pp. 324–344. Springer, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2568–2577. IEEE, 2025.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, et al. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=mSiN7i0BYH.

Yubo Huang, Hailong Guo, Fangtai Wu, Weiqiang Wang, Shifeng Zhang, Shijie Huang, Qijun Gan, Lin Liu, Sirui Zhao, Enhong Chen, et al. Live avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv preprint arXiv:2512.04677*, 2025b.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

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 Transactions on Pattern Analysis and Machine Intelligence*, 2025c.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong MU, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=66NzcRQuOq.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137. IEEE, 2021.

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.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. FIFO-diffusion: Generating infinite videos from text without training. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=uikhNa4wam.

Youngrae Kim, Qixin Hu, C-C Jay Kuo, and Peter Beerel. Memrope: Training-free infinite video generation via evolving memory tokens. In *European Conference on Computer Vision*, pp. 244–261. Springer, 2026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Haodong Li, Shaoteng Liu, Zhe Lin, and Manmohan Chandraker. Rolling sink: Bridging limited-horizon training and open-ended testing in autoregressive video diffusion. *arXiv preprint arXiv:2602.07775*, 2026a.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=X96Ei9n34a.

Mingxiang Liao, Hannan Lu, Qixiang Ye, Wangmeng Zuo, Fang Wan, Tianyu Wang, Yuzhong Zhao, Jingdong Wang, and Xinyu Zhang. Evaluation of text-to-video generation models: A dynamics perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=tmX1AUmkl6.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IAyzXjbfwo.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34385–34397, 2026.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, and Kaipeng Zhang. Yume1. 5: A text-controlled interactive world generation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7752–7761, 2026a.

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. In *European Conference on Computer Vision*, pp. 576–593. Springer, 2026b.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=j7kdXSrISM.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, et al. Matrix-game 3.5: Enhancing real-time streaming interactive world models with patch memory. *arXiv preprint arXiv:2608.29910*, 2026.

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. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=v1DKz5Vxr7.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=j8Vr3E3vhy.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=GfSwkDSr8J.

Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, Linfeng Zhang, et al. Hunyuan-gamecraft-2: Instruction-following interactive game world model. *arXiv preprint arXiv:2511.23429*, 2025.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=P8pqeEkn1H.

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.

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=pYNl76onJL.

Steven Xiao, Xindi Zhang, Dechao Meng, Qi Wang, Peng Zhang, and Bang Zhang. Knot forcing: Taming autoregressive video diffusion models for real-time infinite interactive portrait animation. *arXiv preprint arXiv:2512.21734*, 2025.

Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. LVD-2m: A long-take video dataset with temporally dense captions. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=H5bUdfM55S.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11269–11277, 2026.

Xin Yan, Yuxuan Cai, Qiuyue Wang, Yuan Zhou, Wenhao Huang, and Huan Yang. Long video diffusion generation with segmented cross-attention and content-rich video data curation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3184–3194. IEEE, 2025.

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=sFyTZEqmUY.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=nCAODkpsPJ.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40256–40265, 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. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=gtmyFnvXAW.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Jintao Zhang, Kai Jiang, Jintao Chen, Xu Wang, Deyuan Liu, Jungang Li, Dechuang Chen, Ming Lin, Jingjiang Zhou, Haopeng Jin, et al. Vidu s2: Real-time interactive, editable, and spatial video generation. *arXiv preprint arXiv:2609.11638*, 2026.

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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=J8JCF64aEn.

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.

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. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=BYInOck3gr.

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.

A Extended Related Work

Autoregressive Video Diffusion. Autoregressive video diffusion generates videos by conditioning each chunk on a preceding video prefix. Diffusion Forcing (Chen et al., 2024) combines next-token prediction with sequence diffusion through independent noise levels, while CausVid (Yin et al., 2025) distills a bidirectional teacher into a few-step causal generator. Self Forcing (Huang et al., 2025a) addresses the training–inference gap by training under student self-rollout. Causal Forcing (Zhu et al., 2026) and Causal Forcing++ (Zhao et al., 2026) improve student initialization through causal ODE distillation and causal consistency distillation, respectively. Self Gradient Forcing (Zhuang et al., 2026) uses gradient replay to train memory writing for subsequent generation and evaluates direct initialization from a TF-trained model. Building on direct TF initialization, LongTake studies how supervision from real long videos strengthens the AR model before self-rollout DMD.

Long-Horizon Video Generation. Existing methods extend video generation through sampling and context management. FIFO-Diffusion (Kim et al., 2024) uses diagonal denoising, while StreamingT2V (Henschel et al., 2025) introduces short- and long-term memory. FramePack (Zhang et al., 2025) and PackForcing (Mao et al., 2026b) compress preceding video content, and Deep Forcing (Yi et al., 2026) and Infinity-RoPE (Yesiltepe et al., 2026) adapt cache management and temporal positions during inference. LongLive (Yang et al., 2026), Rolling Forcing (Liu et al., 2026), and MemRoPE (Kim et al., 2026) further improve extended AR rollouts. AdaState (Dalva & Yanardag, 2026) addresses motion stagnation by replacing a fixed anchor with an evolving state. LongTake focuses on learning to predict later frames under conditioning KV-cache states constructed from real long-video prefixes. Such training could complement this line of work by helping the AR model better use the information retained in the KV-cache during long-horizon generation.

Long-Horizon Supervision and Distillation. Long-video data provides supervision for scene evolution beyond short clips. LVD-2M (Xiong et al., 2024) and Presto (Yan et al., 2025) emphasize dynamic long-take videos, while Long Context Tuning (Guo et al., 2025) learns dependencies across extended scenes. Resampling Forcing (Guo et al., 2026) combines longer-video training with robustness to errors in generated video prefixes. For distillation, Self-Forcing++ (Cui et al., 2026) applies DMD to short windows sampled from extended student rollouts using a short-video teacher. MMM (Cai et al., 2026) combines supervised flow matching on long videos with sliding-window distribution matching to a short-video teacher through a shared encoder. Context Forcing (Chen et al., 2026a) trains a long-context teacher on real long videos and applies prefix-conditioned distribution matching. Context-Matched Distillation (Bandyopadhyay et al., 2026) aligns causal teacher scoring with the student self-rollout prefix and uses the teacher to initialize the student. LongTake examines how real long-video supervision improves direct TF initialization under short-horizon joint DMD. The resulting AR teacher can then supervise later frames through Hybrid DMD under student self-rollout.

A.1 Comparison with closely related work.

Long-Horizon TF initialization. Causal Forcing (Zhu et al., 2026) and Causal Forcing++ (Zhao et al., 2026) introduce separate causal ODE and consistency distillation stages before self-rollout DMD. Direct initialization from a TF-trained model is also evaluated in these studies and in SGF (Zhuang et al., 2026). To strengthen this direct initialization, LongTake uses supervision from real long videos. We curate a long-video dataset and fine-tune the AR model through Long-Horizon TF, pairing KV-cache states from long video prefixes with subsequent ground-truth target frames. In Figure 5, we compare short-horizon TF, short-horizon TF followed by causal CD, and Long-Horizon TF under the same subsequent 5s joint DMD configuration. Long-Horizon TF on curated real long videos yields stronger dynamics than short-horizon TF with comparable visual quality, while retaining higher aesthetic quality than causal CD initialization. The resulting two-stage pipeline achieves strong long-horizon generation without a separate few-step initialization stage. These gains show the value of long-video supervision before extending the DMD horizon.

Context Forcing. Context Forcing (Chen et al., 2026a) also curates real long videos to train a long-context AR teacher and uses this teacher for contextual DMD over extended student rollouts. LongTake examines the benefit of this supervision as a direct student initialization while keeping subsequent joint DMD at 5s. The resulting gains show that long-video supervision can improve

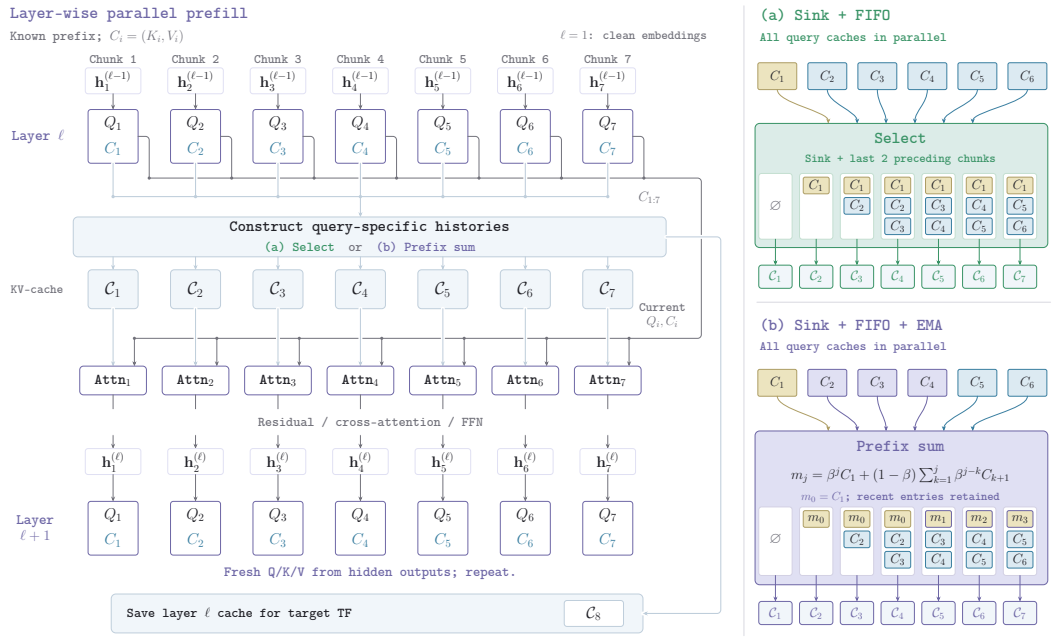

Figure 9: **Parallel cache construction for Long-Horizon TF.** (Left) At each transformer layer, all known prefix chunks are projected together to Q_i and $C_i = (K_i, V_i)$. For each query, the cache operator constructs a conditioning KV-cache $\mathcal{C}_i$, which enters causal attention together with the current chunk's Q/K/V. The resulting hidden states feed the next layer. (Right) Two alternative operators construct the KV-cache states of all queries in parallel. Operator (a) selects the initial sink and the most recent preceding chunks, and operator (b) additionally merges evicted entries into the sink through an associative weighted prefix scan. The example uses seven chunks with one sink and two recent slots. After prefill, $\mathcal{C}_8$ at each layer is saved for the subsequent target-window TF pass. Parallelism is across chunks within each layer, while transformer layers remain sequential.

long-horizon generation before the AR teacher scores extended student rollouts. When distillation is extended through Hybrid DMD, the two methods also differ in their target distributions. Context Forcing matches a target window conditioned on a shared video prefix, whereas our conditional DMD matches the next chunk under the student self-rollout prefix updated before each chunk.

Context-Matched Distillation. Concurrent work, Context-Matched Distillation (Bandyopadhyay et al., 2026), trains a multi-step causal teacher with Diffusion Forcing and directly initializes the student from its weights, without a separate ODE or consistency distillation stage. Its Prefix Scoring conditions teacher scores on a conditioning KV-cache constructed from student-generated video prefixes, while Prefix Corruption stabilizes supervision under imperfect prefixes. LongTake targets the teacher gap with Long-Horizon TF on curated real long videos and shows stronger direct initialization under the same 5s joint DMD procedure. These gains arise before changing the bidirectional scoring or extending the DMD horizon. Hybrid DMD adds AR supervision of later frames while retaining bidirectional joint supervision of the initial window.

B Implementation Details

B.1 Parallel Cache Computation

Long-Horizon TF constructs a conditioning KV-cache from a clean ground-truth video prefix before evaluating the target-window loss. Since all prefix chunks are known, cache construction can process all chunks in parallel within each layer. Figure 9 illustrates this computation for two cache-construction rules: sink retention with FIFO eviction, used in our Long-Horizon TF implementation, and a fixed-coefficient EMA extension.

Shared layer-wise computation. Let the video prefix contain P chunks, and let $\mathbf{h}_i^{(\ell-1)}$ denote the hidden states of chunk i entering transformer layer ℓ. All chunks are projected together to obtain

$$(Q_i^{(\ell)}, K_i^{(\ell)}, V_i^{(\ell)}) = \mathrm{Proj}_\ell(\mathbf{h}_i^{(\ell-1)}), \qquad i = 1, \ldots, P, \tag{8}$$

where Proj_ℓ includes the corresponding normalization and projection operations. We write $C_i^{(\ell)} = (K_i^{(\ell)}, V_i^{(\ell)})$ for one chunk's projected K/V pair, and $\mathcal{C}_i^{(\ell)}$ for the conditioning KV-cache available *before* processing chunk i. The cache operator constructs each $\mathcal{C}_i^{(\ell)}$ from $C_1^{(\ell)}, \ldots, C_{i-1}^{(\ell)}$. Each query then attends to its conditioning KV-cache and its own current K/V,

$$A_i^{(\ell)} = \mathrm{Attn}\Big(Q_i^{(\ell)}; \mathcal{C}_i^{(\ell)} \| C_i^{(\ell)}\Big), \tag{9}$$

where $\|$ concatenates keys and values along the token dimension. Constructing these KV-cache states requires no attention output from another chunk at the same layer, so all query groups can be evaluated in parallel while preserving chunk causality. The remaining transformer operations produce $\mathbf{h}_i^{(\ell)}$, from which layer $\ell+1$ computes fresh projections after layer ℓ completes.

Sink + FIFO via parallel selection. For clarity, consider one sink chunk and a capacity of r recent chunks, with $r = 2$ in Figure 9. Long-Horizon TF uses $r = 5$ with three-frame chunks, which retains three sink frames and 15 recent frames. We omit layer superscripts below. For $i \geq 2$, the recent indices are

$$\mathcal{R}_i = \{j : \max(2, i-r) \leq j \leq i-1\}, \qquad \mathcal{C}_i^{\mathrm{FIFO}} = [C_1] \| [C_j]_{j \in \mathcal{R}_i}, \tag{10}$$

with $\mathcal{C}_1^{\mathrm{FIFO}} = \varnothing$. An empty index interval contributes no entries. The retained indices depend only on the query position and cache capacity, so the KV-cache states of all queries are obtained by parallel indexed selection from the shared K/V bank, without sequentially updating a cache for each query. For example, query 7 uses $[C_1, C_5, C_6]$ when $r = 2$, since C_2 to C_4 have been evicted.

Sink + FIFO + EMA via parallel prefix scan. The EMA extension additionally merges evicted entries into the sink with a fixed coefficient $\beta \in [0, 1)$ (Lu et al., 2026). With equal-sized chunks,

$$m_0 = C_1, \qquad m_j = \beta m_{j-1} + (1-\beta) C_{j+1} = \beta^j C_1 + (1-\beta) \textstyle\sum_{k=1}^{j} \beta^{j-k} C_{k+1}. \tag{11}$$

The same weights are applied separately to keys and values. Before query $i \geq 2$, exactly $e_i = \max(0, i-r-2)$ non-sink chunks have been evicted, giving

$$\mathcal{C}_i^{\mathrm{EMA}} = [m_{e_i}] \| [C_j]_{j \in \mathcal{R}_i}, \qquad \mathcal{C}_1^{\mathrm{EMA}} = \varnothing. \tag{12}$$

Although the recurrence is written sequentially, its updates are affine maps $m \mapsto am + b$ whose ordered composition is associative,

$$(a_2, b_2) \circ (a_1, b_1) = (a_2 a_1,\ a_2 b_1 + b_2). \tag{13}$$

An associative prefix scan (Blelloch, 1990) therefore computes all required sink states in parallel with logarithmic scan depth. Each query selects its corresponding sink state and recent entries. For example, query 7 uses $[m_3, C_5, C_6]$ when $r = 2$, where m_3 merges C_2 to C_4 into the sink.

Saved cache and target prediction. At each layer, the same operator constructs $\mathcal{C}_{P+1}^{(\ell)}$, the cache after all P prefix chunks. For the seven-chunk example, this is $[C_1, C_6, C_7]$ for FIFO or $[m_4, C_6, C_7]$ for EMA. Collecting these final caches across layers yields the conditioning KV-cache $\mathtt{c}_{\mathtt{gt}}^{s}$ used in Algorithm 1. For matching query-specific cache selection, aggregation, and positional treatment, the layer-wise schedule is mathematically equivalent to sequential prefix processing; numerical differences can arise from floating-point accumulation order. The prefix pass runs without gradient tracking, and the saved cache is detached. A second, two-stream forward pass evaluates TF losses over the fixed target window: clean chunks provide preceding ground-truth context, while noisy chunks predict target velocities. Gradients remain enabled through the clean-context computation within this target window.

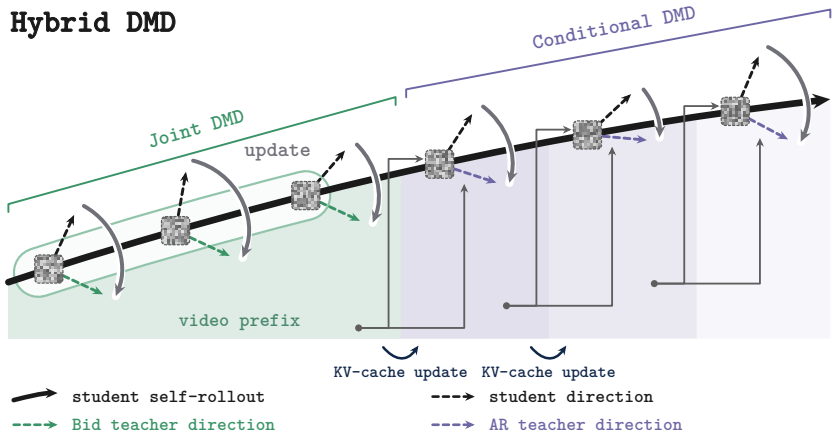

Figure 10: **Hybrid DMD.** A bidirectional teacher jointly supervises the initial window. The Long-Horizon AR teacher supervises later frames using the conditioning KV-cache updated during student self-rollout for each target chunk.

Bounded context within the target window. The saved prefix cache supplies the initial history for target prediction. For each target chunk, the same sink-retention and FIFO rule is applied to the saved history and preceding clean target chunks in temporal order. The original sink is retained, while the recent slots contain at most the five most recent non-sink chunks. Thus, each query uses at most 18 past latent frames plus its current three-frame chunk, giving a maximum attention window of 21 latent frames. The clean and noisy query streams use the same past-history selection, but each supplies its own current K/V. In particular, a noisy query cannot access the clean K/V of its own target chunk or any future chunk. These query-specific selections implement the cache updates within the packed two-stream forward pass.

Bounded positional encoding. In our FIFO implementation, keys are stored before the rotary positional transformation and selected separately for each query. If a query retains $h \leq 18$ past latent frames, the concatenated history and current keys receive temporal RoPE positions $0, \ldots, h+2$, while the current queries receive positions $h, h+1, h+2$. At full capacity, the sink occupies positions 0–2, the recent frames occupy 3–17, and the current chunk occupies 18–20. This reindexing keeps the temporal positions within the 21-frame range regardless of the target's original position in the long video. For $s = 0$, no prefix pass is needed and the computation reduces to standard TF.

B.2 Hybrid DMD

We implement the joint objective in (4) using Self Gradient Forcing (Zhuang et al., 2026). We describe how gradient replay computes the student updates and how these updates extend to Hybrid DMD. Model configurations and training settings are provided in Appendix C.1.

Student rollout and gradient replay. We first perform student self-rollout without gradient tracking. On each device, we uniformly select one denoising step and share this selection across all chunks. Although the rollout completes all denoising steps, we record the noisy inputs only at the selected step, together with the corresponding inputs for encoding the video prefix.

We then replay these recorded inputs through a packed two-stream student forward pass. The prefix stream encodes the recorded student self-rollout prefix, while the noisy stream predicts each current chunk. Causal masking prevents the noisy stream from accessing current or future target frames in the prefix stream. Let $\hat{\mathbf{x}}_\theta^{<N}$ denote the resulting clean predictions over the initial joint window.

Although both recorded input streams are detached, replay recomputes the video-prefix representations, allowing gradients to reach the student parameters that encode the prefix. The update therefore differentiates through replay without backpropagating through the preceding student self-rollout.

Hybrid DMD update. We write the two contributions to the objective in (7) as $\mathcal{J}_{\texttt{DMD}}^{\texttt{Bi}}$ and $\mathcal{J}_{\texttt{DMD}}^{\texttt{AR}}$. The student update combines these contributions over the initial window and later frames as

$$\widehat{\nabla}_\theta \mathcal{J}_{\texttt{Hybrid}} = \widehat{\nabla}_\theta \mathcal{J}_{\texttt{DMD}}^{\texttt{Bi}} + \lambda \widehat{\nabla}_\theta \mathcal{J}_{\texttt{DMD}}^{\texttt{AR}}. \tag{14}$$

Here, $\widehat{\nabla}_\theta$ denotes the update estimated using detached self-rollout inputs and SGF replay. The following expressions describe the joint and conditional contributions in terms of score differences.

Joint contribution. Let $s_\psi^{\texttt{Bi}}$ denote the trainable joint fake-score model and $s^{\texttt{Bi}}$ the frozen bidirectional teacher score. Both models evaluate the complete initial noisy window $\mathbf{x}_t^{<N}$ at the same timestep. The joint contribution follows from their score difference over this window

$$\widehat{\nabla}_\theta \mathcal{J}_{\texttt{DMD}}^{\texttt{Bi}} = \mathbb{E}\left[w(t) \left(s_\psi^{\texttt{Bi}}(t, \mathbf{x}_t^{<N}) - s^{\texttt{Bi}}(t, \mathbf{x}_t^{<N}) \right)^\top \frac{\partial \mathbf{x}_t^{<N}}{\partial \theta} \right]. \tag{15}$$

The expectation is over student self-rollout, the sampled denoising step for replay, and the score-evaluation timestep and noise. Although generation is autoregressive, both score models evaluate the initial window jointly rather than scoring each target chunk under a separately updated prefix.

Conditional contribution. For each target chunk beyond the initial window, we evaluate the conditional fake-score model $s_\psi^{\texttt{AR}}$ and the frozen Long-Horizon AR teacher $s_{\phi^\star}^{\texttt{AR}}$ under the same student self-rollout prefix. Using half-open temporal intervals, the corresponding chunk indices satisfy $N \le i < M$. Averaging over these target chunks gives the conditional contribution

$$\widehat{\nabla}_\theta \mathcal{J}_{\texttt{DMD}}^{\texttt{AR}} = \frac{1}{M-N} \sum_{i=N}^{M-1} \mathbb{E}\left[w(t_i) \left(s_\psi^{\texttt{AR}}\left(t_i, \mathbf{x}_{t_i}^i, \mathtt{c}_{\texttt{self}}^i\right) - s_{\phi^\star}^{\texttt{AR}}\left(t_i, \mathbf{x}_{t_i}^i, \mathtt{c}_{\texttt{self}}^i\right) \right)^\top \frac{\partial \mathbf{x}_{t_i}^i}{\partial \theta} \right]. \tag{16}$$

Each expectation additionally includes the sampled student self-rollout prefix. The cache notation identifies the shared conditioning video prefix, which each score model encodes using its own parameters. Noise is applied only to the target chunk, so both conditional scores use an unperturbed student self-rollout prefix while evaluating the same noisy target chunk.

We treat the sampled student self-rollout prefix as fixed when computing this update. Thus, gradients do not propagate through the generation of preceding chunks, but student replay retains gradients through the video-prefix representations recomputed from this fixed prefix. Averaging over target chunks beyond the initial window then gives the conditional update used during training.

C Experimental Details

C.1 Training Configuration

Model initialization. We initialize Long-Horizon TF from the released chunk-wise AR diffusion checkpoint of Causal Forcing[2] (Zhu et al., 2026), based on `Wan2.1-T2V-1.3B` (Wan et al., 2025). After 3,000 additional updates, the resulting model initializes the student and serves as the frozen Long-Horizon AR teacher. LongTake then trains the student with joint DMD over 21 latent frames for 1,200 iterations using Self Gradient Forcing (Zhuang et al., 2026). For Hybrid DMD, training branches after 800 joint DMD iterations and proceeds on 42-latent-frame rollouts for another 400 iterations. In these extended rollouts, a frozen `Wan2.1-T2V-14B` teacher jointly supervises the first 21 latent frames, while the Long-Horizon AR teacher conditionally supervises the remaining 21.

Optimization. Table 3 summarizes the AdamW (Loshchilov & Hutter, 2019) configuration for both training stages. During distillation, we update the fake-score models every iteration and the student every five iterations. We initialize parameter EMA after iteration 200 and update it after each subsequent student update with decay 0.99. For each variant, we use the same final EMA student and KV-cache size for both 30s and 60s evaluation, keeping model parameters and cache capacity fixed.

[2] https://huggingface.co/zhuhz22/Causal-Forcing/blob/main/chunkwise/ar_diffusion.pt under Apache-2.0

Table 3: **Training configuration.** LongTake uses 1,200 joint DMD iterations. The Hybrid DMD variant uses 800 joint DMD iterations followed by 400 Hybrid DMD iterations.

Setting	Long-Horizon TF	Joint DMD	Hybrid DMD
Iterations	3,000	1,200 / 800	400
Generator learning rate	5×10^{-7}	2×10^{-6}	2×10^{-6}
Fake-score learning rate	—	4×10^{-7}	4×10^{-7}
Effective batch size	64	8	8
Gradient accumulation	8	1	1
Optimizer	AdamW	AdamW	AdamW
(β_1, β_2)	$(0, 0.999)$	$(0, 0.999)$	$(0, 0.999)$
Weight decay	0.01	0.01	0.01
Learning-rate schedule	Constant	Constant	Constant
Gradient clipping	10	10	10
Parameter EMA decay	—	0.99	0.99

Table 4: Prefix sampling for Long-Horizon TF. Prefix lengths are measured in latent frames and sampled in multiples of three. Probabilities refer to the complete training mixture.

Source	Prefix length s	Probability
Causal Forcing clean GT	0	0.10
OpenVidHD	0–15	0.20
OpenVidHD	18–39	0.45
OpenVidHD	42–60	0.20
OpenVidHD	63–99	0.05

Initialization comparison. In Figure 5, Short TF, Short TF + CD, and Long TF use the same 5s joint DMD configuration in Table 3, differing only in student initialization. Short TF and Short TF + CD start from the short-horizon AR teacher and the causal CD checkpoint released by Causal Forcing (Zhu et al., 2026; Zhao et al., 2026), respectively. For Short TF, we evaluate the released SGF EMA checkpoint (Zhuang et al., 2026) trained with this configuration, while Long TF starts from our Long-Horizon TF model trained on the curated long-video dataset. We evaluate all three models on 30s rollouts using the same protocol, so the comparison measures the effect of initialization under a shared distillation procedure.

Long-Horizon TF. Each update supervises a target window of 21 latent frames following a ground-truth video prefix. We sample the prefix length s in multiples of three according to Table 4. After selecting a bin and sampling a valid prefix length uniformly within that bin, we sample a compatible video length in proportion to its number of clips and select eight distinct clips for the eight devices. Although clips are distinct within each microbatch, sampling across microbatches uses replacement. Each clip contains at least $s + 21$ latent frames, and the maximum prefix length is 99, so supervision reaches 120 latent frames with the loss window fixed at 21.

For $s > 0$, we construct the conditioning KV-cache with one parallel pass through the transformer without gradient tracking. A second pass then jointly processes the clean context and noisy target streams. The conditioning KV-cache is detached, while gradients are retained through the clean-context computation within the target window. For $s = 0$, no prefix cache is needed, so standard TF processes the clean context and noisy target together in a single transformer pass.

Cache configuration. Long-Horizon TF, the AR teacher, and the conditional fake-score model retain three sink frames and 15 recent frames. Together with the current three-frame chunk, these form an attention window of 21 latent frames. Following (Zhuang et al., 2026), the student uses three sink frames and six recent frames, giving a 12-frame attention window including the current chunk.

Distillation and sampling. The four-step student uses denoising timesteps $t \in \{1, 0.9375, 0.8333, 0.625\}$. Using the guidance convention $u + \omega(c - u)$, the bidirectional and AR teachers use $\omega = 4$ and $\omega = 3$, respectively. For Hybrid DMD, the generator loss is the mean joint surrogate plus $\lambda = 0.2$ times the mean conditional surrogate over the seven target chunks

covering the later frames. Both fake-score models use flow regression on fresh student samples, with equal weights for their two losses. For the positive-λ comparisons, runs branch from the same state after 800 joint-DMD iterations and receive 400 Hybrid-DMD iterations. The $\lambda = 0$ control uses joint DMD for all 1,200 iterations, keeping the total distillation budget fixed across Table 2.

C.2 Training Data

Video corpus. Long-Horizon TF uses 43,000 video–text pairs comprising 40,000 OpenVidHD clips[3] (Nan et al., 2025) and 3,000 short clips from the released Causal Forcing clean-synthesized data[4] (Zhu et al., 2026). OpenVidHD clips contain 21–120 latent frames, corresponding to approximately 5–30 seconds at 16 FPS. The short subset contains 21 latent frames per clip, and we retain the original captions and clean-GT prompts for both sources. During training, we sample OpenVidHD with probability 0.9 and the short subset with probability 0.1.

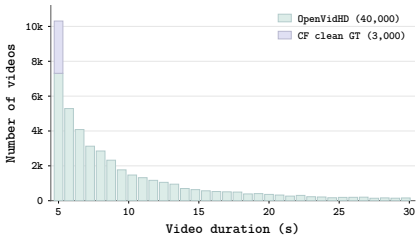

Figure 11: **Dataset duration.**

OpenVidHD curation. We begin with 433,509 registry entries from the OpenVidHD (Nan et al., 2025) subset of OpenVid-1M. We require a source frame rate of at least 16 FPS and a candidate latent length between 21 and 123, computed from the source video metadata as

$$\left\lfloor \frac{16\,(\text{source frame count})/(\text{source FPS}) + 3}{4} \right\rfloor . \tag{17}$$

These criteria retain 243,010 candidates. Since OpenVidHD clips generally exhibit limited dynamics, we select 40,000 videos from these candidates to match the motion statistics of the short-horizon TF reference data while filtering for visual quality and temporal consistency.

C.3 Baseline Configuration

Baselines. Self Forcing (Huang et al., 2025a) uses the released DMD EMA checkpoint. Causal Forcing (Zhu et al., 2026) uses its released chunk-wise DMD generator. LongLive (Yang et al., 2026) uses the released base model with its rank-256 LoRA and a 12-frame window containing three sink frames, six recent frames, and the current chunk. Reward Forcing (Lu et al., 2026) uses a nine-frame window containing three sink frames, three recent frames, and the current chunk. It updates the sink with weights 0.999 on the previous sink and 0.001 on evicted content. MemRoPE (Kim et al., 2026) uses 12-frame window contains three sink frames, two EMA memory frames, four recent frames, and the current chunk, with new-content weights 0.01 and 0.1 for the two memory frames. Rolling Forcing† denotes the official long-video model released by Causal Forcing (Zhu et al., 2026), which combines the Rolling Forcing framework (Liu et al., 2026) with causal ODE initialization. We use this released four-step model with its 24-frame rolling-window configuration. SGF (Zhuang et al., 2026) uses the released chunk-wise TF-initialized EMA checkpoint with three sink frames, six recent frames, and the current chunk. Context Forcing (Chen et al., 2026a) uses its released EMA checkpoint with native Slow-Fast Memory. This configuration retains three sink frames within a 21-frame cache and selects keyframes from the video prefix by similarity.

C.4 Evaluation Protocol

Prompts and generation. The 30s evaluation uses the first 128 prompts from the extended MovieGen prompt set (Polyak et al., 2024) distributed with Self Forcing (Huang et al., 2025a). For 60s evaluation, we use the union of the six VBench-Long (Huang et al., 2025c) quality-dimension prompt sets, containing 251 unique prompts. Within each evaluation horizon, all methods use the same prompts. Table 6 lists the number of prompts evaluated for each video-quality dimension.

[3] https://huggingface.co/datasets/nkp37/OpenVid-1M under CC BY 4.0
[4] https://huggingface.co/zhuhz22/Causal-Forcing-data

Table 5: **Full evaluation on 30s and 60s autoregressive rollouts.** `Quality` aggregates seven video-quality dimensions using the standard VBench normalization and weights. LongTake uses Long-Horizon TF and joint DMD on 5s rollouts. The best and second-best results are highlighted.

	30s							
Method	Subj.↑	Backg.↑	Motion↑	Dyn.↑	Aesth.↑	Imaging↑	Flicker↑	Quality↑
Short-video supervision								
Self Forcing (Huang et al., 2025a)	97.43	96.50	98.55	40.42	60.66	69.27	98.86	82.34
Causal Forcing (Zhu et al., 2026)	96.10	95.76	96.99	69.38	57.41	66.66	98.24	82.19
LongLive (Yang et al., 2026)	98.03	97.01	98.78	41.25	63.21	69.50	99.06	83.25
Reward Forcing (Lu et al., 2026)	97.26	96.49	98.15	66.98	62.42	69.64	98.61	84.36
MemRoPE (Kim et al., 2026)	97.28	96.50	98.19	58.65	62.33	70.62	98.77	83.95
Rolling Forcing† (Liu et al., 2026)	96.04	95.44	97.48	93.44	58.61	69.08	97.96	84.67
Self Gradient Forcing (Zhuang et al., 2026)	98.19	97.02	98.75	51.30	63.90	71.55	98.56	84.25
Long-video supervision								
Context Forcing (Chen et al., 2026a)	97.72	96.86	98.47	57.71	63.05	68.73	99.14	84.16
LongTake (Ours)	96.83	96.11	97.38	90.73	62.51	70.68	98.59	85.79
+ Hybrid DMD ($\lambda = 0.2$)	96.32	95.62	97.26	96.30	62.17	70.65	98.75	85.97
	60s							
Method	Subj.↑	Backg.↑	Motion↑	Dyn.↑	Aesth.↑	Imaging↑	Flicker↑	Quality↑
Short-video supervision								
Self Forcing (Huang et al., 2025a)	97.27	96.33	98.66	44.21	56.58	67.87	99.13	81.89
Causal Forcing (Zhu et al., 2026)	95.96	96.17	97.52	59.86	51.83	65.15	98.77	80.93
LongLive (Yang et al., 2026)	98.40	97.10	98.80	37.04	62.23	69.48	99.32	82.98
Reward Forcing (Lu et al., 2026)	97.29	96.13	98.45	80.23	62.08	69.95	98.79	85.54
MemRoPE (Kim et al., 2026)	97.48	96.33	98.69	56.20	59.34	72.20	99.03	83.92
Rolling Forcing† (Liu et al., 2026)	95.81	95.14	97.63	95.69	56.75	68.67	98.72	84.78
Self Gradient Forcing (Zhuang et al., 2026)	98.42	96.47	98.77	53.06	64.09	69.40	98.87	84.15
Long-video supervision								
Context Forcing (Chen et al., 2026a)	98.17	96.75	99.00	52.69	62.10	67.37	99.39	83.85
LongTake (Ours)	96.78	95.52	97.64	94.72	62.35	69.91	98.60	85.96
+ Hybrid DMD ($\lambda = 0.2$)	96.43	94.57	97.47	97.36	61.48	70.17	98.57	85.71

Table 6: Evaluation prompt counts.

Dimension	30s	60s
Subject consistency	128	72
Background consistency	128	86
Motion smoothness	128	72
Dynamic degree	128	72
Aesthetic quality	128	93
Imaging quality	128	93
Temporal flickering candidates	128	75

IQ Drift. Following (Liu et al., 2026), we measure endpoint image-quality changes using the VBench MUSIQ evaluator (Ke et al., 2021) with the SPAQ checkpoint (Fang et al., 2020). For each video, we average frame-level scores over the first and last five seconds, using all 80 frames in each window. Let $m(\mathbf{x}_{i,t})$ denote the MUSIQ score on its original point scale and let $W_{\texttt{start}}$ and $W_{\texttt{end}}$ denote these windows. We report the mean absolute difference between the two endpoint scores as

$$\Delta\texttt{IQ} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{1}{80}\sum_{t\in W_{\texttt{end}}} m(\mathbf{x}_{i,t}) - \frac{1}{80}\sum_{t\in W_{\texttt{start}}} m(\mathbf{x}_{i,t})\right|. \tag{18}$$

We take the absolute difference before averaging over the 128 MovieGen videos at 30s or the 93-video imaging-quality subset at 60s. Windows are extracted losslessly from the original RGB videos without further temporal splitting. We use the longer preprocessing mode, which resizes 832×480 frames to 512×295 and scales pixel values by $1/255$. Lower $\Delta\texttt{IQ}$ indicates smaller endpoint quality changes. However, consistently low-quality videos and videos with intermediate degradation can

Table 7: **VLM** exposure-stability levels, summarized from the original MemRoPE rubric.

Score	Exposure criterion
0	Near-total whiteout or blackout makes the scene unreadable.
1	Widespread exposure failure severely reduces visibility.
2	Persistent highlight or shadow clipping causes substantial detail loss.
3	Exposure problems have limited spatial extent or duration.
4	Occasional local exposure flaws cause little visibility loss.
5	Balanced exposure preserves visibility without disruptive clipping or darkening.

also obtain low scores, so we interpret Δ**IQ** together with imaging quality. To distinguish endpoint degradation from improvement, we additionally retain the signed end-minus-start difference, whose negative mean indicates degradation in endpoint image quality on average.

VLM Exposure. We use the exposure-stability rubric of MemRoPE (Cui et al., 2026; Kim et al., 2026) to evaluate overexposure, underexposure, and the resulting loss of visibility. We use Gemini-3.1-pro-preview (model identifier `gemini-3.1-pro-preview`) as the judge. Each request contains one full original 30s or 60s video and the unchanged exposure rubric, without the text prompt used for generation. We fix the API settings to 1 FPS analysis sampling, high media resolution, temperature 1.0, low thinking level, and a maximum of 4,096 output tokens. The original video remains at 16 FPS and is uploaded without re-encoding or temporal splitting. The judge returns an integer score $e_i \in \{0, 1, 2, 3, 4, 5\}$ and a textual justification. Table 7 summarizes the scoring levels, where higher scores indicate better exposure quality and stability. We report the mean over all 128 MovieGen prompts (Polyak et al., 2024) for 30s generation and all 251 VBench-Long prompts (Huang et al., 2024) for 60s generation. Failed requests are retried without assigning a score, while successful judgments are not repeated for score selection. Each reported mean includes valid judgments for every video in its prompt set under the same unchanged rubric and fixed API settings.

Aggregate score. We compute the Quality score using the standard VBench (Huang et al., 2024) normalization and weighting of the seven video-quality dimensions. Dynamic degree receives a weight of 0.5, while each remaining dimension receives a weight of 1, and we scale the weighted mean by 100. Separately, Δ**IQ** and **VLM** Exposure assess visual stability and do not contribute to the Quality score.

D Additional Results

D.1 Teacher Generation Before Distillation

We compare Short TF and Long TF before DMD by generating subsequent frames from the same real video prefix. We use 50 held-out real long videos that were not used for Long-Horizon TF fine-tuning. For each video, each teacher constructs its conditioning KV-cache from approximately 5s or 20s of real video (81 or 321 video frames, respectively) and then generates approximately 5s of additional video (81 video frames), yielding 200 outputs in total. Each teacher updates its conditioning KV-cache from generated frames, retaining 3 sink frames and 15 recent latent frames.

Evaluation. We evaluate 24 uniformly sampled generated frames, excluding the real video prefix. VisionReward (**VR**) (Xu et al., 2026) aggregates 29 binary video judgments using the official weights. Instruction following (**IF**) evaluates whether the generated video satisfies some requirements of the prompt. Responses are encoded as $+1$ for yes and -1 for no and averaged across videos. Both scores are scaled by 100, with **IF** reported as a signed average. We assess the immediate transition from observed to generated frames using boundary **DINO** feature similarity (Caron et al., 2021) and normalized RGB L1 difference between the last observed frame and the first generated frame.

Results. In Table 8, the Long-Horizon TF model trained on curated real long videos yields improvements in evaluated generation quality at both prefix lengths. The boundary metrics suggest better preservation of visual appearance at the transition from observed to generated frames, although these measurements are limited to the immediate transition and can favor static outputs. Because **IF** scores

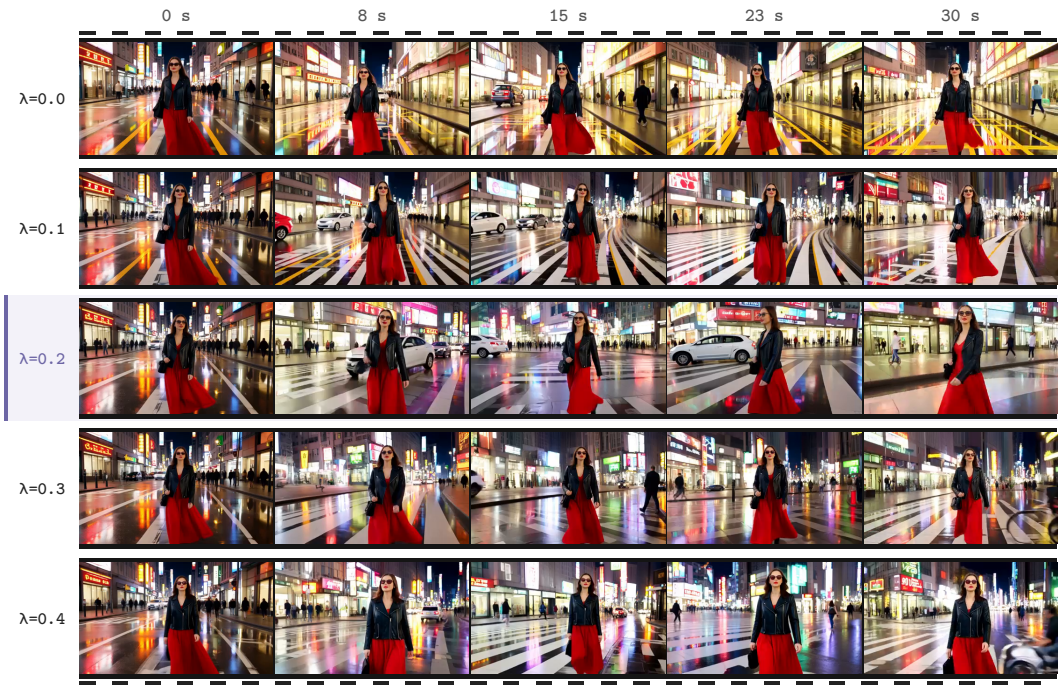

Figure 12: **Effect of the conditional DMD weight λ.** Frames sampled from 30s rollouts show how λ affects scene progression. With $\lambda = 0.2$, both the viewpoint and surrounding scene evolve while preserving a coherent appearance of the central subject throughout the rollout.

Table 8: **Teacher generation before DMD on held-out real long videos.** Each teacher constructs its conditioning KV-cache from the same real video prefix and generates 81 additional video frames. Quality and prompt alignment are evaluated on generated frames, while boundary metrics assess appearance preservation at the immediate transition from observed to generated video frames.

Prefix (video frames)	Teacher	VR ↑	IF ↑	DINO ↑	RGB L1 ↓
81 ($\approx 5s$)	Short TF	17.26	96	0.9904	0.0202
	Long TF	18.28	100	0.9919	0.0176
321($\approx 20s$)	Short TF	16.60	96	0.9930	0.0183
	Long TF	17.83	100	0.9941	0.0166

are near saturation, they provide limited evidence of a difference in prompt alignment between the two teachers.

D.2 Motion-Quality Trade-off

Figure 13 extends Figure 8 to 60s rollouts. Both LongTake variants remain on the Pareto front of dynamic degree and aesthetic quality, and the same holds when aesthetic quality is replaced by non-motion quality, the VBench Quality score recomputed without dynamic degree (Figure 14).

D.3 Additional Qualitative Results

We complement the quantitative evaluation with additional visual comparisons of long-horizon generation, TF initialization, and conditional supervision through Hybrid DMD.

Effect of the conditional DMD weight. Figure 12 compares 30s rollouts with different values of λ. In this example, $\lambda = 0.2$ produces more pronounced scene progression than $\lambda = 0$, while preserving a coherent appearance of the subject throughout the generated sequence.

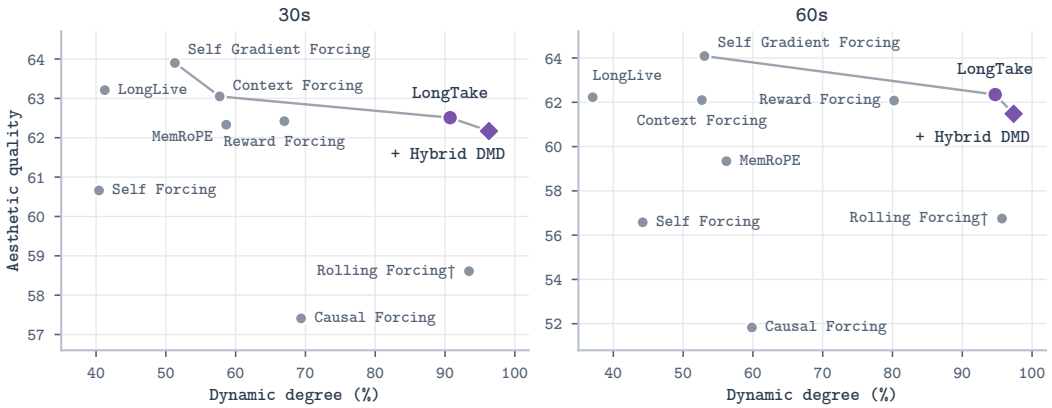

Figure 13: **Motion-quality trade-off at (Left) 30s and (Right) 60s.** The gray line joins the Pareto-optimal methods. No other method has both higher dynamic degree and higher aesthetic quality. Both LongTake variants lie on front at both durations.

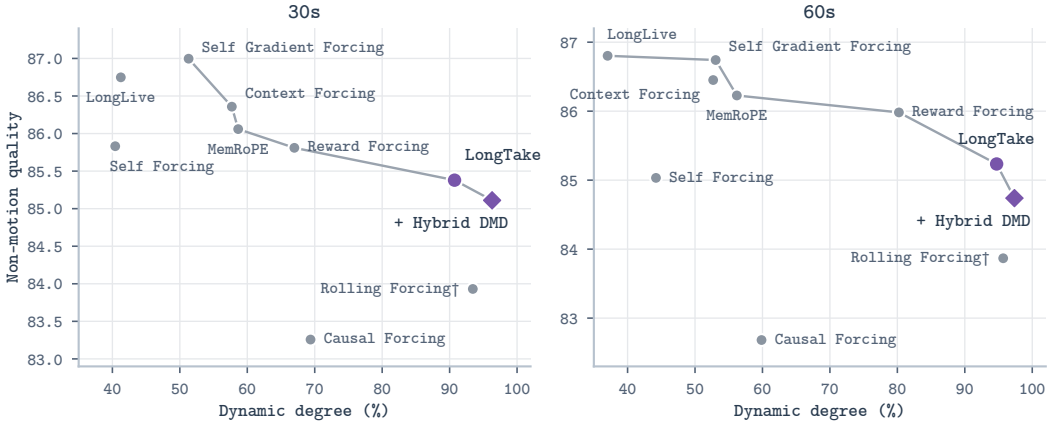

Figure 14: **Motion vs. non-motion quality** Non-motion quality is the VBench Quality score recomputed without dynamic degree with equal weights. Both LongTake variants lie on the front at both durations and are the only methods on it with dynamic degree above 90.

Long-horizon generation. Figures 15 to 17 provide additional baseline comparisons on 30s and 60s generation, complementing the examples in Figure 7 with different subjects and scenes.

Effect of Long-Horizon TF. Figures 18 to 20 provide additional comparisons illustrating the effect of Long-Horizon TF initialization under the same 5s joint DMD procedure.

Effect of Hybrid DMD. Figures 21 and 22 provide additional comparisons illustrating the effect of Hybrid DMD. These examples complement Table 2, where all settings use the same Long-Horizon TF initialization and are evaluated after 1,200 distillation iterations. Compared with joint DMD ($\lambda = 0$), Hybrid DMD ($\lambda = 0.2$) increases dynamics, with modest reductions in subject consistency and aesthetic quality. With the teacher-training stage held fixed, this comparison shows the additional motion–quality tradeoff introduced by extending supervision through Hybrid DMD.

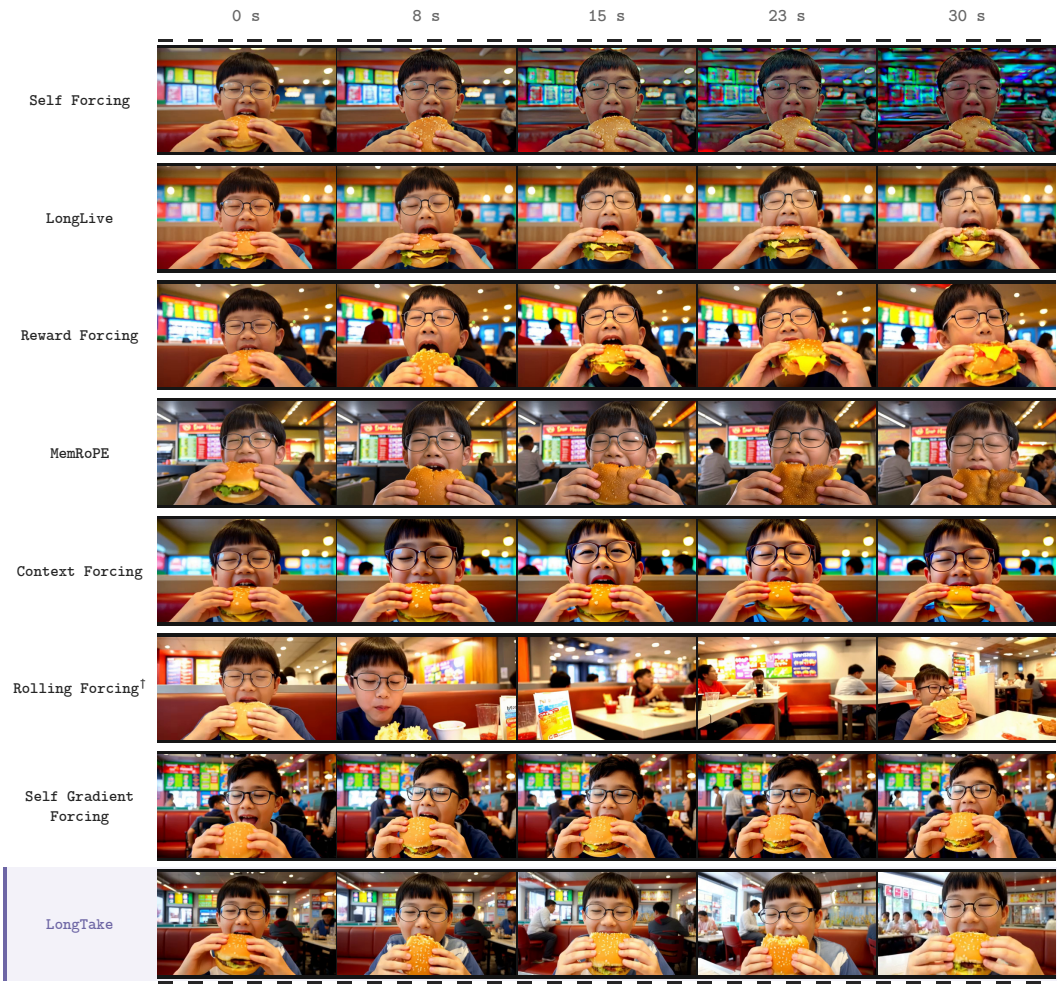

Figure 15: **30s generation with the same prompt across evaluated methods.**

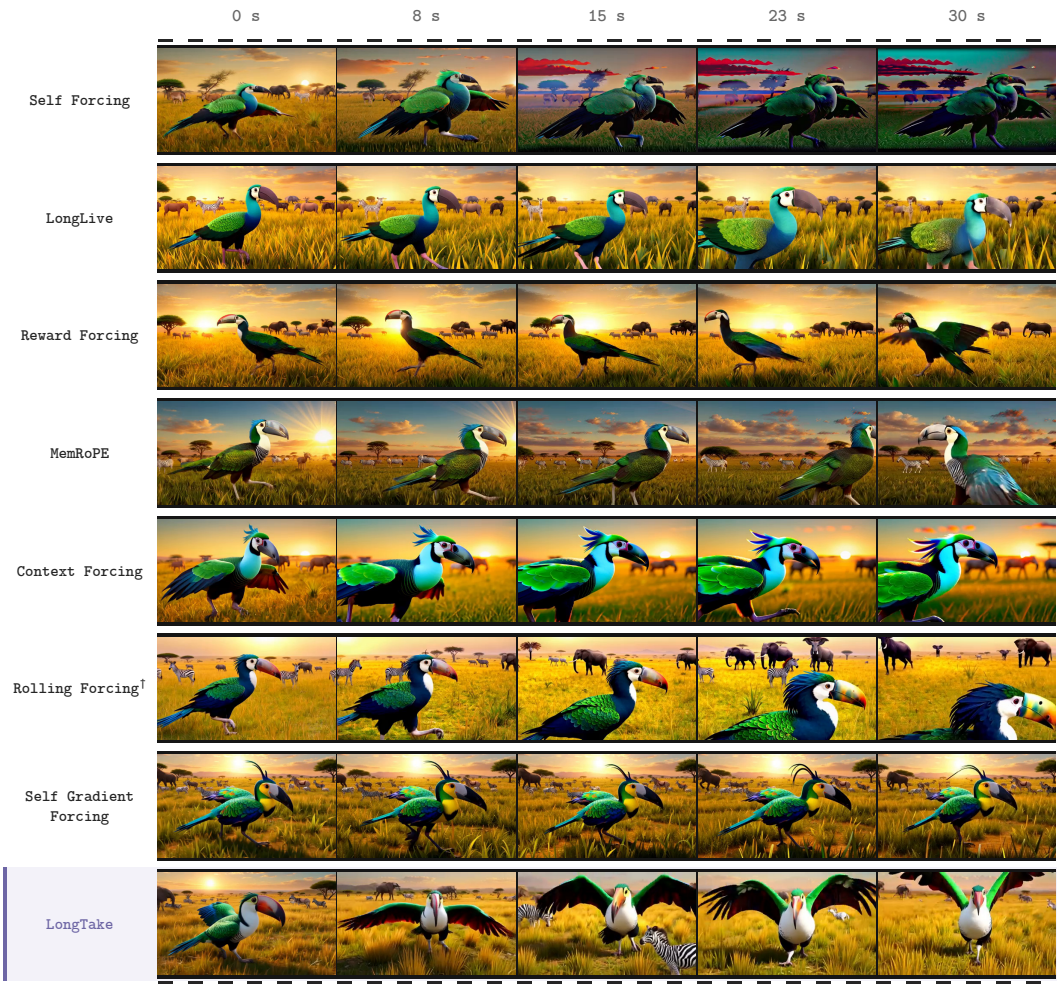

Figure 16: **30s generation with the same prompt across evaluated methods.**



Figure 17: **60s generation with the same prompt across evaluated methods.**



Figure 18: **Long-Horizon TF initialization under the same 5s joint DMD.**

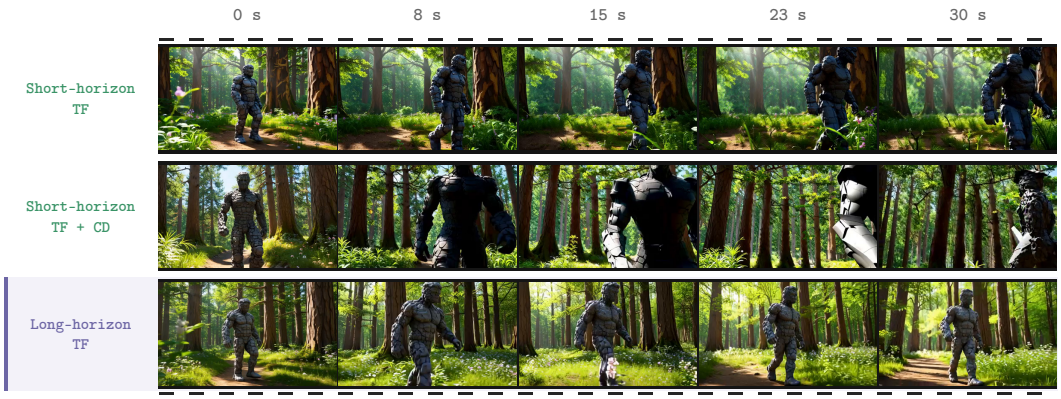

Figure 19: **Long-Horizon TF initialization under the same 5s joint DMD.**

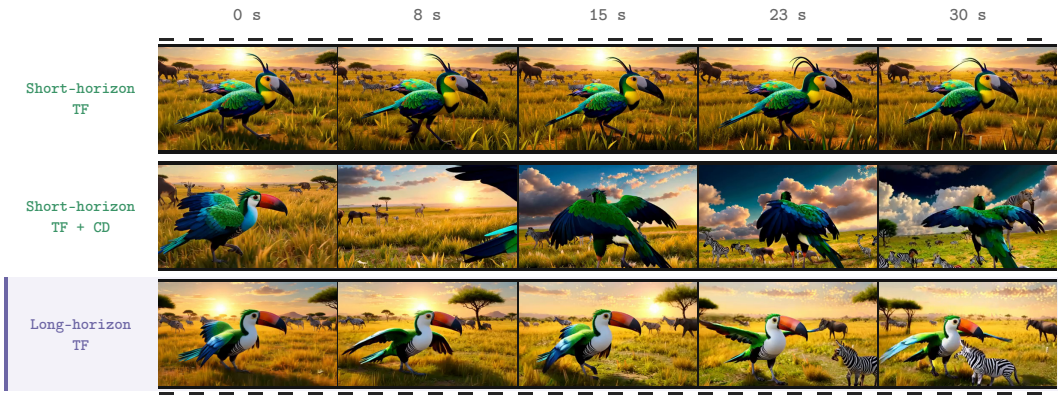

Figure 20: **Long-Horizon TF initialization under the same 5s joint DMD.**

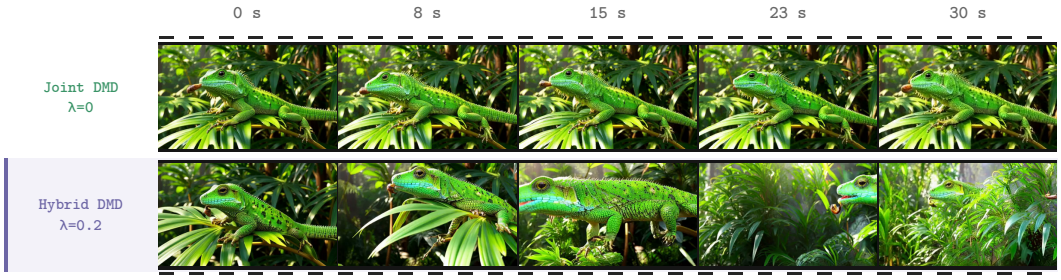

Figure 21: **Hybrid DMD with shared initialization and equal distillation iterations.**



Figure 22: **Hybrid DMD with shared initialization and equal distillation iterations.**