Anchor Forcing: Anchor Memory and Tri-Region RoPE for Interactive Streaming Video Diffusion

Yang Yang[1,2], Tianyi Zhang[2], Wei Huang[3], Jinwei Chen[2], Boxi Wu[1⋆], Xiaofei He[1], Deng Cai[1], Bo Li[2], and Peng-Tao Jiang[2⋆]

[1] Zhejiang University, China
[2] vivo BlueImage Lab, vivo Mobile Communication Co., Ltd., China
[3] University of Hong Kong, China
https://vivocameraresearch.github.io/anchorforcing/

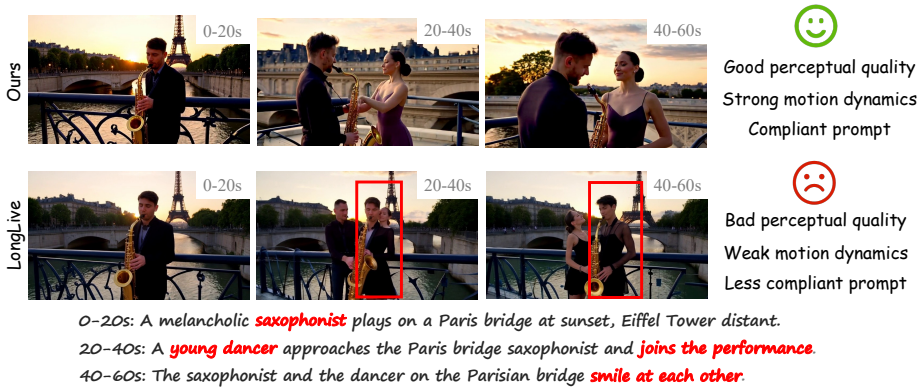

Fig. 1: We propose **Anchor Forcing**, which supports prompt switches that introduce new subjects and actions while preserving context, motion quality, and temporal coherence across clips. In contrast, prior methods degrade over time and often fail to realize newly introduced interactions, as highlighted by the red boxes. Red text denotes the interaction newly specified in each segment.

Abstract. Interactive long video generation requires prompt switching to introduce new subjects or events, while maintaining perceptual fidelity and coherent motion over extended horizons. Recent distilled streaming video diffusion models reuse a rolling KV cache for long-range generation, enabling prompt-switch interaction through re-cache at each switch. However, existing streaming methods still exhibit progressive quality degradation and weakened motion dynamics. We identify two failure modes specific to interactive streaming generation: (i) at each prompt switch, current cache maintenance cannot simultaneously retain KV-based semantic context and recent latent cues, resulting in weak

⋆ Corresponding author.

boundary conditioning and reduced perceptual quality; and (ii) during distillation, unbounded time indexing induces a positional distribution shift from the pretrained backbone's bounded RoPE regime, weakening pretrained motion priors and long-horizon motion retention. To address these issues, we propose **Anchor Forcing**, a cache-centric framework with two designs. First, an anchor-guided re-cache mechanism stores KV states in anchor caches and warm-starts re-cache from these anchors at each prompt switch, reducing post-switch evidence loss and stabilizing perceptual quality. Second, a tri-region RoPE with region-specific reference origins, together with RoPE re-alignment distillation, reconciles unbounded streaming indices with the pretrained RoPE regime to better retain motion priors. Experiments on long videos show that our method improves perceptual quality and motion metrics over prior streaming baselines in interactive settings.

Keywords: Interactive Video Generation · Quality Preservation · Motion Enhancement

1 Introduction

Interactive long video generation is becoming increasingly important for creative content creation [27,48] and film-level scene development [1,23]. Compared with short clips, long-horizon generation must sustain narrative coherence while continuously producing visually faithful frames and plausible temporal dynamics. A straightforward approach [1, 21, 23, 27, 48] that applies dense spatiotemporal attention over all frames is computationally prohibitive at long horizons, which makes high-quality generation difficult to deploy in interactive scenarios where users expect low latency and frequent edits.

To improve practicality, recent distilled streaming video diffusion models [18, 41–43] adopt causal temporal attention and reuse a rolling transformer KV cache. With causal attention [47], each denoising step conditions only on previously generated frames, while cached keys and values enable efficient reuse of long-range context across steps without recomputing attention over the entire history. In practice, these systems typically combine short-window attention with a small set of sink tokens [40] to stabilize long-horizon generation and support prompt-switch interaction by updating the cache state at switch boundaries.

Despite this progress, existing approaches largely emphasize extending temporal length, while interactive long video generation still exhibits two recurring failure modes. As shown in Fig. 1, cache maintenance at prompt switches often fails to preserve relevant semantic evidence, weakening cross-segment continuity in interactive videos and hurting perceptual quality. In particular, re-cache-based methods [18, 41] rebuild the cache by discarding prior semantic KV states from earlier stages, which reduces the available semantic context for post-switch conditioning and degrades perceptual quality in interactive long videos. Second, during distillation, streaming generation [18, 41] uses RoPE indices that grow unbounded over time, while the pretrained backbone is optimized under a finite and bounded positional range. This gap induces a positional distribution

shift, weakening the transfer and retention of pretrained motion priors. Together, these effects make interactive long videos prone to degraded visual quality and weakened motion dynamics.

To bridge these gaps, we propose Anchor Forcing, a cache-centric framework for interactive long video generation under streaming constraints, designed to preserve visual quality and retain motion dynamics. Our framework has two core designs. First, we introduce an anchor-guided re-cache mechanism. At each prompt switch, we generate a short initial segment conditioned on the new prompt and store its key–value states as a junction cache. We then build an anchor memory by integrating the junction cache with the sink cache collected at the early stage of generation and the local cache that preserves the most recent context. This anchor memory is then used to warm-start re-caching for subsequent generations under the new prompt, reducing evidence loss and stabilizing post-switch quality. Second, we propose a tri-region RoPE with region-specific reference origins and indexing. Instead of using an unbounded positional index, we adopt a relative, range-bounded indexing scheme that keeps effective RoPE positions within the bounded regime observed during pretraining. Moreover, we treat the sink, junction, and local caches separately and assign each region its own RoPE encoding. During distillation, the generator adaptively attends across regions and leverages complementary cues, improving motion-prior retention and long-horizon dynamics.

Empirically, Anchor Forcing improves perceptual quality and motion dynamics on VBench benchmarks. In summary, our contributions are:

- We introduce an anchor cache that stores anchor memory after each prompt switch, enabling anchor-guided re-cache to reduce post-switch evidence loss and stabilize generation quality.
- We propose a tri-region RoPE that partitions the anchor memory into three regions, keeps effective RoPE positions within the bounded range seen in pretraining, and encourages adaptive cross-region feature retrieval during distillation, thereby improving long-horizon motion-prior retention.
- We achieve state-of-the-art results on VBench in interactive settings on perceptual quality and motion-related metrics, while consistently outperforming prior streaming baselines overall.

2 Related Work

2.1 Diffusion-Based Long Video Generation

A straightforward route to longer videos is to scale the context length of video diffusion Transformers and train models to utilize large spatiotemporal windows [1, 4, 6, 10, 12, 23, 27, 28, 34]. For example, SVI [23] enables infinite-length video generation by error-recycling fine-tuning that trains the model to detect and correct its own accumulated errors. FreeNoise [28] adopts noise rescheduling with window-based temporal attention. LVDM [12] performs hierarchical video synthesis in a compact 3D latent space to improve efficiency. Mixture of

Contexts [1] learns sparse routing to a few informative segments plus mandatory anchors, improving efficiency compared to uniform long-context attention. Seine [4] uses random-mask conditioned video diffusion to extend shot-level clips into story-level videos with text-guided transitions. SSM [27] investigates state-space based architectures to extend temporal memory without quadratic attention growth. Dalal et al. [6] augment a pretrained diffusion transformer with test-time training (TTT) layers, maintaining long-horizon consistency while keeping self-attention computationally tractable. LaVie [37] adopts a cascaded generation pipeline to enhance long-range coherence. LCT [10] expands pretrained single-shot video diffusion models to multi-shot scenes by enlarging the context window and extending full attention across shots. Despite these advances, bidirectional processing and repeated denoising steps still pose practical bottlenecks for streaming and prompt-switchable generation.

2.2 Autoregressive-Based Long Video Generation

To enable real-time or long-horizon streaming, a growing line of work [2, 5, 7, 9, 13, 15, 18, 19, 29, 32, 34, 41–44, 47] adopts causal temporal modeling and reuses temporal KV caches. CausVid [47] reframes video diffusion as a causal generation process and reduces the number of inference steps via distribution-matching distillation. [45]. Self-Forcing [15] reduces the train–test mismatch in autoregressive video diffusion by simulating KV caching during training, thereby mitigating exposure bias. Infinity-RoPE [42] studies position-encoding and cache behaviors under long rollouts and proposes KV flush to improve horizon and responsiveness during interaction. LongLive [41] supports prompt switching in long rollouts via frame-level autoregressive generation with short-window causal attention, frame-sink tokens, and a KV-recache update scheme. MemFlow [18] enables narrative prompting by using prompt-conditioned memory retrieval with sparse activation over a memory bank. Nevertheless, streaming methods still exhibit long-horizon failure modes, including drift and regressions toward early content. LoL [5]analyzes sink-collapse behavior and proposes a training-free mitigation to reduce such degeneracy at extreme horizons. LongVie [8] introduces multimodal guidance, unified noise initialization, and degradation-aware training, and StreamingT2V [13] extends these ideas with short- and long-term memory modules for coherent text-to-video generation. SkyReels-V2 [3] combines diffusion forcing with a film-structure planner and multimodal controls. FramePack [49] compresses input frames into a fixed-size context to address memory and efficiency constraints, and FAR [9] further improves autoregressive generation by coupling a high-resolution short-term context with a compressed long-term context via flexible positional encoding. History-guided video diffusion [29] incorporates flexible-length historical context to strengthen temporal consistency over extended rollouts, while Pyramidal-flow [19] proposes a multi-scale flow-matching design to reduce computation. In addition, recent training-free horizon extension methods [20, 22, 24–26, 31, 33, 39, 50] aim to extend pretrained or short-horizon-trained models to longer temporal ranges without full model retraining.

Overall, existing autoregressive streaming approaches substantially improve efficiency and temporal extent. Yet, two issues remain: cache maintenance at prompt switches fails to jointly retain KV-based semantic context and recent latent cues, weakening boundary conditioning and perceptual quality, and unbounded time indexing during distillation deviates from the pretrained backbone's bounded RoPE regime, undermining motion-prior transfer and long-horizon motion retention.

3 Method

3.1 Preliminaries

Streaming autoregressive video diffusion. We consider streaming autoregressive (AR) video diffusion that generates long videos in small temporal chunks while reusing a temporal Key–Value (KV) cache to reduce per-step computation [47]. At each step, the model conditions on (i) a local cache that stores recent latent tokens to capture short-term dynamics, (ii) a small global sink cache that provides a persistent surrogate of long-range context, and (iii) the current text prompt embedding. Practical systems [18, 41] typically combine short-window causal attention with a bounded KV cache composed of local and sink caches, enabling long-horizon generation with bounded attention context. Moreover, they often employ distribution matching distillation [45] to distill multi-step diffusion models into a few-step generator. Given a noised latent x_t at diffusion timestep t and prompt embedding c, the model predicts a denoising direction conditioned on the assembled memory:

$$\hat{v}_\theta = f_\theta(x_t,\, t,\, c;\, \mathcal{M}), \tag{1}$$

where $\mathcal{M}$ denotes the memory constructed from the sink cache and the rolling local cache, and is updated in a rolling manner after each generated chunk. Interaction is realized by updating the conditioning prompt at interaction boundaries during inference (including prompt switches and edits), requiring the model to incorporate updated instructions while maintaining temporal coherence.

Distribution Matching Distillation. Let G_θ denote the few-step generator parameterized by θ, which maps Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ to a generated sample $x_0 = G_\theta(\epsilon)$. To distill a multi-step diffusion model into G_θ, the distribution matching distillation [45, 46] (DMD) minimizes the reverse KL divergence between the real data distribution p_{data} and the output distribution of the generator p_{gen}. The gradient of the reverse KL divergence can be approximated as the difference between two score functions:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = \mathbb{E}_t \left[\nabla_\theta \text{KL}(p_{\text{gen}}(x_t) || p_{\text{data}}(x_t))\right] \approx \mathbb{E}_{x_t,t}\left[\left(s_{\text{gen}}(x_t) - s_{\text{data}}(x_t)\right) \frac{\partial x_t}{\partial \theta}\right], \tag{2}$$

where $x_t = \alpha_t x_0 + \beta_t \epsilon'$ denotes the re-noised sample of x_0 using the noise schedule coefficients α_t and β_t with $\epsilon' \sim \mathcal{N}(0, I)$, s_{data} and s_{gen} represent the score functions of the real data and generator's output distribution, respectively. During

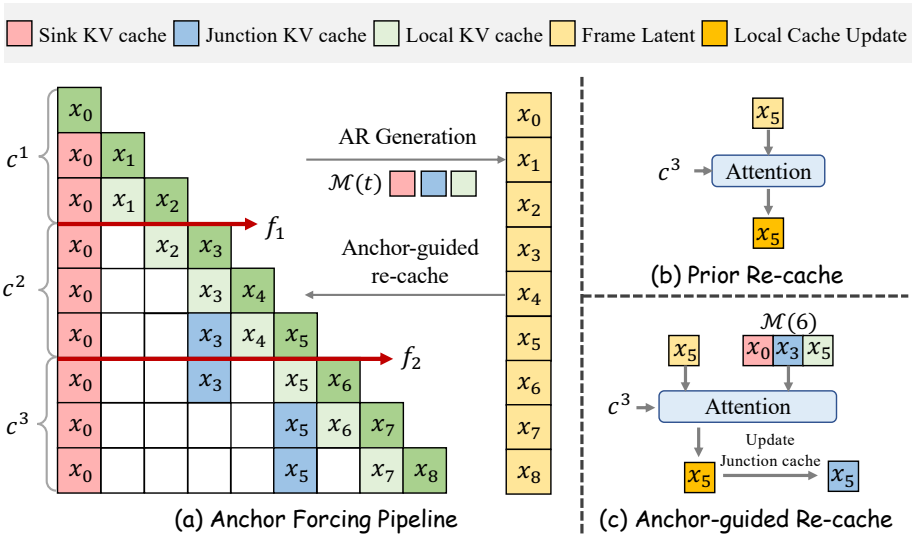

Fig. 2: The Anchor Forcing pipeline. (a) **Overview of Anchor Forcing** in an interactive setting with two prompt switches. We denote the anchor memory for generating frame t as $\mathcal{M}(t)$, and apply anchor-guided re-cache at f_1 and f_2 to update the local KV cache under the new prompt condition. (b) **Prior re-cache [41]**. It rebuilds the local cache solely from historical frame latents, which fails to retain prior KV evidence across prompt switches. (c) **Anchor-guided re-cache** at f_2. It augments re-cache with the anchor memory $\mathcal{M}(6)$ and refreshes the junction caches x_5.

training, DMD uses a fixed pre-trained teacher diffusion model to approximate the score function of the real data distribution, and simultaneously trains a neural network to estimate the score function of the generator's output distribution. Finally, the gradient derived in Eq. 2 is used to guide the generator to align its output distribution with the data distribution.

3.2 Anchor-guided Re-cache Mechanism

Problem formulation. In interactive text-to-video generation, the conditioning prompt is updated at interaction boundaries, and a cache maintenance operation is typically invoked to rebuild the KV cache for the new prompt condition. As illustrated in Fig. 2(b). Specifically, re-cache-based methods [18, 41] rebuild the cache by evicting prior KV states, which removes informative context from the preceding stage and weakens cross-switch semantic continuity. Overall, these deficiencies reduce the availability of effective conditioning cues around interaction boundaries, leading to degraded video quality in interactive long videos.

Anchor Memory. We maintain an anchor memory $\mathcal{M}(t)$ composed of three cache parts. The sink cache $\mathcal{S}$ stores KV states from the earliest N_S generated frames and remains fixed throughout generation, providing global semantic

context. The junction cache stores boundary-adjacent evidence captured immediately after each prompt switch, preserving newly introduced interaction cues. The local cache retains the most recent short-range context to support fine-grained continuity. We consider a prompt stream $\mathcal{C} = \{c^1, c^2, \ldots, c^m\}$ with interaction boundaries $\mathcal{F} = \{f_1, f_2, \ldots, f_{m-1}\}$, where f_i denotes the frame index at which the prompt updates to c^i. For any frame t, we define:

$$\pi(t) = \max\{\, i \mid f_i \le t,\ i \in [1, m-1] \,\}, \qquad s(t) = f_{\pi(t)}. \tag{3}$$

The active prompt at frame t is $c^{\pi(t)}$, and $s(t)$ is the most recent frame index of the interaction boundary. We denote by $\mathbf{KV}[a:b]$ the per-layer KV state of frames a to b for causal Transformer layers [47]. After each boundary at f_i, we take the first N_J frames under the updated prompt and store their per-layer KV states in the caches:

$$\mathcal{J}_i = \mathbf{KV}[f_i : f_i + N_J - 1]\,. \tag{4}$$

The junction caches are refreshed at each interaction boundary and kept fixed until the next boundary. In contrast, the rolling local KV cache maintains only a window of length W:

$$\mathcal{M}_\ell(t) = \mathbf{KV}[t - W + 1 : t]\,. \tag{5}$$

For simplicity, as shown in Fig. 2(a), we illustrate the rolling update of the anchor memory at the two prompt switches, f_1 and f_2, under the setting $N_J = W = 1$. To avoid redundancy with local KV cache, junction caches are activated only after the junction frames have left the local window. Accordingly, the anchor memory $\mathcal{M}(t)$ used to generate frame t is

$$\mathcal{M}(t) = \big[\mathcal{S}\ ;\ \mathcal{M}_\ell(t)\ ;\ \delta_t\, \mathcal{J}_{\pi(t)}\big],\ \text{where } \delta_t = \mathbb{I}[s(t) + N_J \le t - W]\,. \tag{6}$$

When $\delta_t = 0$, the junction frames are still contained in $\mathcal{M}_\ell(t)$ and their KV states are directly accessible via the local context. When $\delta_t = 1$, they have been evicted by rolling updates, and $\mathcal{J}_{\pi(t)}$ provides a persistent substitute that keeps junction evidence reachable and repeatedly attendable in later segments.

Anchor-guided Re-cache. In interactive text-to-video generation, the prompt is updated at interaction boundaries, and a re-cache operation is typically invoked to rebuild the cache state under the new prompt condition. However, as shown in Fig. 2(b), re-cache [41] discards prior KV states, which reduces the available semantic context for post-switch conditioning and removes boundary-relevant evidence that is crucial for continuity and high-fidelity generation after the switch. To mitigate this issue, we propose an anchor-guided re-cache mechanism, as illustrated in Fig. 2(c). Unlike prior re-cache, we recompute the local cache under the updated prompt by additionally leveraging the anchor memory from the previous segment. Subsequent streaming generation is conditioned on the refreshed local cache together with the sink cache, with the junction cache providing persistent boundary-adjacent evidence to support smooth motion evolution while promptly adapting to the updated prompt.

3.3 Tri-region RoPE

RoPE [30] is widely used to encode temporal positions in Transformer attention. In streaming generation [18, 41] with rolling KV caches, a straightforward implementation assigns RoPE indices using a global latent frame index that grows monotonically over time. As the rollout length increases, these indices may exceed the positional range observed during pretraining, inducing a positional distribution shift and weakening the transfer of pretrained motion priors and long-horizon motion retention.

In our setting, the generator is initialized from CausVid [47], which is trained with ODE trajectories capped at 21 latent frames (distilled from Wan [35]); thus, the model is primarily exposed to RoPE indices within a bounded range. Accordingly, during streaming inference we replace global absolute indexing with relative, range-bounded indexing when attending to cached tokens, and cap the effective RoPE index to $P_{\max} = 21$ to stay within the pretrained positional range. We adopt a relative, range-bounded indexing scheme, termed tri-region RoPE. Specifically, when generating frame t, the anchor memory is $\mathcal{M}(t) = [\mathcal{S}\ ;\ \mathcal{M}_\ell(t)\ ;\ \delta_t \mathcal{J}_{\pi(t)}]$. We then assign region-specific, bounded RoPE indices to cached keys from the sink, local, and clip regions as follows. Sink caches form a small persistent memory with a fixed order. We use 0 to $N_S - 1$ as its index. Let the local cache cover a rolling window ending at t. When $t < P_{\max}$, we use the latent index as the RoPE index. When $t \geq P_{\max}$, we re-index keys relative to the $P_{\max}$ and window of length W:

$$p^{\mathrm{L}}(k,t) = \begin{cases} t, & t < P_{\max}, \\ P_{\max} - W + k, & t \geq P_{\max}, \end{cases} \quad where \quad k \in [0, W-1] \tag{7}$$

where k denotes the index of the local caches. At the same time, the query of the current frame t also uses the same encoding form. When the junction cache $\mathcal{J}_{\pi(t)}$ is activated, we place its indices immediately in front of the local cache:

$$p^{\mathrm{J}}(k,t) = p^{\mathrm{L}}(k,t) - N_J. \tag{8}$$

These definitions keep the RoPE indices used in attention bounded by $P_{\max}$, rather than growing with the rollout length. Coupled with distillation training, the generator can better exploit the induced relative geometry across cache regions, learning to retrieve complementary semantic cues from different regions, which leads to stronger motion dynamics in long-horizon interactive generation.

4 Experiments

4.1 Implementation Details

Anchor Forcing is built upon Wan2.1-T2V-1.3B [35] and generates 5-second videos at 832×480 resolution. We first train the model using 16k ODE sampled from the base model, and initialize it with causal attention masking following

Table 1: Interactive long video evaluation results. The dynamic degree and quality scores are reported on the whole 60s sequence. CLIP scores are reported on 10s video segments with the same semantics (↑ higher is better).

Method	Dynamic Degree ↑	Quality Score ↑	CLIP Score ↑					
			0–10s	10–20s	20–30s	30–40s	40–50s	50–60s
Infinity-RoPE [42]	35.17	79.98	23.87	22.62	22.16	21.82	22.20	22.20
MemFlow [18]	45.47	79.35	**24.87**	24.08	23.02	22.87	22.35	21.59
LongLive [41]	26.37	78.92	24.82	24.18	23.30	22.98	23.40	22.82
Anchor Forcing	**73.00**	**82.25**	24.48	**24.35**	**23.50**	**23.53**	**23.61**	**23.63**

CausVid [47]. Text prompts are drawn from the filtered and LLM-augmented VidProM [36] dataset. We then follow LongLive [41] to train the model with DMD, switch to short-window attention with sink tokens, and perform streaming long-tuning. During streaming long-tuning, we use LoRA [14] with rank 256, where each iteration continues the model's own rollout by generating consecutive 5-second clips up to a maximum length of 60 seconds. In our implementation, the sink caches and junction caches store KV states for $N_S = 3$ and $N_J = 3$ latent frames, respectively, while the rolling local caches use a window size of $W = 9$. We optimize both the actor and critic using AdamW, with learning rates $lr = 1.0 \times 10^{-5}$ and $lr_{\text{critic}} = 2.0 \times 10^{-6}$. Training takes approximately 24 hours on 56 A800 GPUs.

4.2 Metrics

We evaluate our method using VBench [16] and VBenchLong [17]. For the 60-second interactive setting and all of the ablation study, we evaluate VBench-Long [17] dimensions including subject consistency, background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality. Following the standard VBench protocol [16], these dimensions are normalized and aggregated using the official coefficients to obtain the overall score. For semantic adherence under prompt interaction, we segment each video at prompt boundaries and compute a clip-wise semantic score using ViCLIP [38]. For the 30-second and 5-second single-prompt settings, we follow Self-Forcing [15] by using their rewritten prompts, and evaluate all VBench dimensions.

4.3 Comparisons for Interactive Generation

For interactive long-form videos with multiple prompt switches, few prior methods support true streaming generation. We implement this setting for three representative baselines, namely Infinity-RoPE [42], MemFlow [18], and LongLive [41], and compare our approach against them. For a fair comparison, we follow the interactive benchmark introduced in MemFlow [18], which contains 100 narrative scripts. Each script comprises six consecutive 10-second prompts, producing

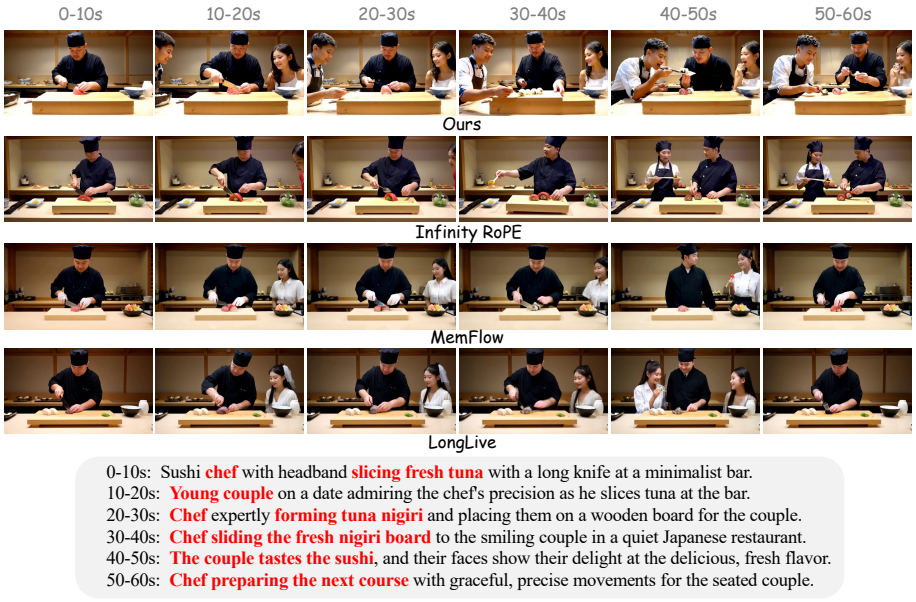

Fig. 3: Qualitative comparison on interactive long-video generation. Compared with baselines, Anchor Forcing achieves stronger prompt compliance, more coherent dynamic motion, and higher long-range visual quality. Red text marks the interactive content newly introduced in each segment.

100 videos with a total duration of 60 seconds. We evaluate dynamic degree and overall quality on the full 60-second sequence using VBenchLong [17], and report segment-wise semantic adherence by computing ViCLIP [38] scores on each 10-second segment with its corresponding prompt.

As shown in Tab. 1, Anchor Forcing achieves the best dynamic degree and quality score among all methods, demonstrating strong motion dynamics and perceptual quality in interactive long videos. Moreover, while early-segment semantic scores are comparable across methods, Anchor Forcing maintains higher CLIP scores in later segments with substantially less temporal degradation, indicating more stable semantic adherence over long horizons under prompt interaction. Additionally, Fig. 3 presents a challenging multi-stage interactive scenario and a qualitative comparison with representative streaming baselines. Our method follows prompt-induced interactions more faithfully: it introduces the new subjects in the 10–20s segment as specified and responds to subsequent action prompts in the 30–60s segment with more accurate and temporally coherent motion. Compared with prior methods, our results exhibit stronger motion dynamics, with smoother transitions and more consistent content updates across prompt switches, while maintaining higher visual fidelity throughout the sequence. In contrast, representative baselines often struggle to follow interactive prompts and sustain strong dynamics: they fail to reliably introduce the "young

Table 2: Comparison with relevant baselines under 5s setting. We compare Anchor Forcing with representative open-source models of similar parameter sizes and resolutions. Evaluation scores are calculated on the standard prompt of VBench.

Model	#Params	Resolution	Dynamic ↑	Evaluation scores ↑		
				Total	Quality	Semantic
Real-data-trained models						
Pyramid Flow [19]	2B	640×384	68.06	81.56	83.89	72.24
LTX-Video [11]	1.9B	768×512	75.00	82.33	85.20	70.83
SkyReels-V2 [3]	1.3B	960×540	48.61	81.72	82.93	76.90
Wan2.1 [35]	1.3B	832×480	73.61	84.27	85.24	80.43
Streaming AR models						
CausVid [47]	1.3B	832×480	63.89	82.96	83.94	79.02
Self-Forcing [15]	1.3B	832×480	63.89	83.83	84.60	80.79
Infinity-RoPE [42]	1.3B	832×480	58.33	83.55	84.34	80.43
DeepForcing [43]	1.3B	832×480	63.89	83.85	84.61	80.84
MemFlow [18]	1.3B	832×480	50.00	81.76	83.74	73.86
LongLive [41]	1.3B	832×480	37.50	82.81	83.26	**81.01**
Anchor Forcing	1.3B	832×480	**73.61**	**83.99**	**84.84**	80.57

couple" in 10–20s and produce weak or ambiguous actions (e.g., "slides a board of fresh nigiri" and "tastes the sushi") in 30–60s, whereas Anchor Forcing generates clearer, more plausible prompt-aligned motion.

4.4 Comparisons for Non-Interactive Generation

To assess the generality of Anchor Forcing beyond prompt interaction, we further compare it with representative non-interactive streaming autoregressive video generation models under single-prompt settings at 5s and 30s. All methods are evaluated on the official VBench prompt set with 946 prompts, using the same resolution of 832×480 and the same Wan2.1 [35] backbone.

We compare Anchor Forcing with CausVid [47], Self-Forcing [15], Infinity-RoPE [42], DeepForcing [43], MemFlow [18], and LongLive [41]. These streaming AR baselines are distilled from Wan-based teachers and share similar model sizes and resolutions. As shown in Tab. 2, on the 5-second benchmark, Anchor Forcing achieves the highest overall VBench score among streaming autoregressive models, delivering the best perceptual quality and strong motion performance. Moreover, our method preserves more of the teacher model's capabilities, particularly in dynamics, achieving performance comparable to the teacher. Long-horizon single-prompt generation is more challenging due to accumulated distribution shift and compounding temporal errors. We therefore evaluate 30s generation using VBenchLong metrics and report total, quality, semantic, and dynamic scores. As shown in Tab. 3, for 30s single-prompt generation, Anchor Forcing ranks first overall and achieves the best total and quality scores. This improvement is consistent across sub-metrics, with the highest imaging quality and temporal style,

Table 3: Comparison with relevant baselines under 30s setting. We compare Anchor Forcing with representative open-source models. Evaluation scores are calculated on the standard prompt of VBench-Long.

Model	Total Score ↑	Quality Score ↑	Semantic Score ↑	Dynamic Score ↑	Imaging Quality ↑	Temporal Style ↑
Self-Forcing [15]	82.21	83.38	77.54	51.85	68.17	23.31
Infinity-RoPE [42]	83.16	84.17	79.13	57.04	67.94	23.79
DeepForcing [43]	82.26	83.20	78.50	52.86	67.65	23.51
MemFlow [18]	82.95	83.86	79.31	60.46	67.68	23.89
LongLive [41]	82.77	83.31	**80.64**	40.19	68.96	23.97
Anchor Forcing	**83.25**	**84.40**	78.67	**79.17**	**69.37**	**24.17**

which indicates stronger long-horizon visual fidelity and stylistic stability. We further provide qualitative results in Fig. 4. Compared with prior methods, Anchor Forcing produces higher-quality renderings and follows the prompt more faithfully. In contrast, other baselines exhibit noticeable degradations in either the background or the main content, highlighted by the yellow and red boxes. For example, DeepForcing, MemFlow, and Self-Forcing show progressive background corruption over time, manifesting as ring-like artifacts. Infinity-RoPE and LongLive suffer from content inconsistencies between 15s and 25s, where the number of paper-origami dancers fluctuates across adjacent clips.

4.5 User study

To better assess the practical quality of generated videos, we conduct a human study under the same interactive setting. For each method, we use the same random seed to generate a one-minute video consisting of six prompt-driven interaction segments. Participants are shown paired results from our method and baseline methods, and are asked to choose their preferred video from three aspects: temporal coherence, prompt following, and perceptual quality. In total, the study involves 20 participants and 660 pairwise preference choices. As shown in Tab. 4, our method is preferred by a clear margin across all evaluated aspects, achieving 64.55% preference in temporal coherence, 67.73% in prompt following, and 63.18% in perceptual quality. In contrast, prior streaming baselines receive substantially lower preference scores on all three dimensions. These results indicate that Anchor Forcing not only improves automatic metrics but also produces videos that are more temporally coherent, better aligned with interactive prompts, and visually more appealing to human observers.

4.6 Ablation Studies

Impact of Tri-region RoPE. Adding tri-region RoPE (tr-RoPE) to the baseline yields the largest improvement in motion. As shown in Tab. 5, the dynamic

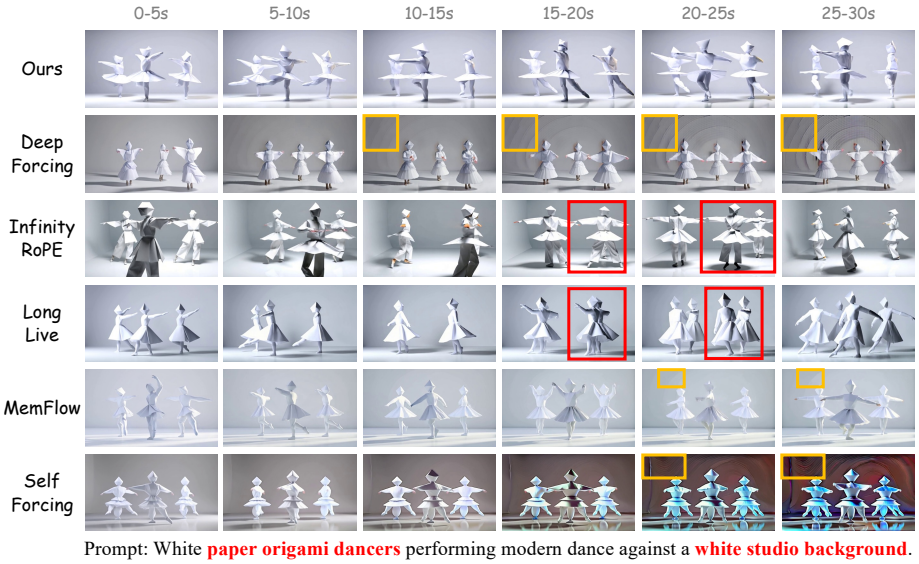

Fig. 4: Qualitative comparison on 30-second long-video generation. Compared with prior methods, ours method yields higher-quality, more prompt-faithful videos, avoiding the background and content degradations highlighted in the yellow and red boxes.

Table 4: Human preference comparison of different methods.

Method	Temporal Coherence↑	Prompt Following↑	Perceptual Quality↑
LongLive	15.45%	15.00%	14.55%
Infinity-RoPE	10.00%	7.27%	11.82%
MemFlow	10.00%	10.00%	10.45%
Ours	**64.55%**	**67.73%**	**63.18%**

Table 5: Ablation study on key components. The best results for the module are indicated in bold.

Method	CLIP Score ↑	Dynamic Degree ↑	Quality Score ↑
Baseline	23.58	26.37	78.92
+ tr-RoPE	23.71	**74.83**	81.72
+ tr-RoPE + AR	**23.85**	73.00	**82.25**

degree increases from 26.37 to 74.83, indicating that tr-RoPE effectively enhances long-horizon dynamics by using region-wise, range-bounded indexing that keeps RoPE positions within the pretrained regime. We further provide qualitative comparisons in the second row of Fig. 5: tr-RoPE produces noticeably more coherent motion than the baseline, but still exhibits appearance degradation and inconsistent prompt responsiveness in some interactive cases. For example, in the highlighted regions, the meat skewer details degrade during 0–20s, and the tourist attributes become less faithful during 40–60s. This observation motivates introducing junction caches and anchor-guided re-cache to better preserve boundary evidence and stabilize visual quality across prompt switches.

Impact of Anchor-guided Re-cache Mechanism. On top of tr-RoPE, introducing the anchor-guided re-cache mechanism (AR) further improves perceptual quality and text alignment. As shown in the last row of Tab. 5, adding AR in-

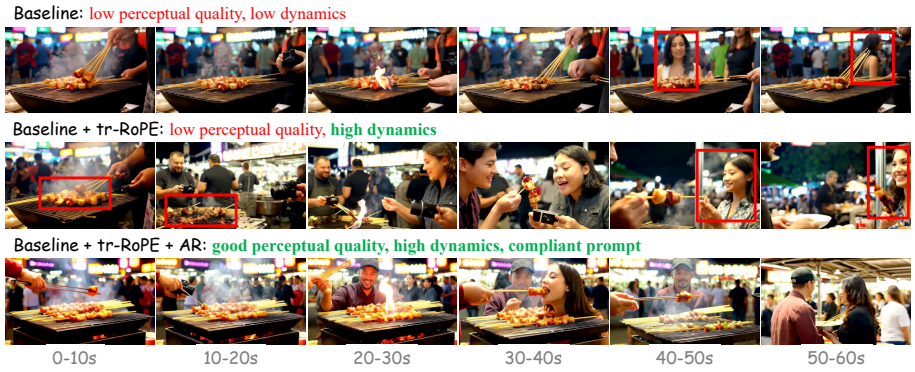

Fig. 5: Ablation on 60-second interactive generation. Tri-region RoPE improves motion dynamics, and anchor-guided re-cache further preserves perceptual quality and prompt compliance across segments. Red boxes mark degradation between adjacent clips, and red text denotes newly introduced interactions.

creases the quality score from 81.72 to 82.25 and the CLIP score from 23.71 to 23.85, indicating that AR improves video quality and stabilizes content across prompt switches. We also provide a qualitative comparison in a 60s interactive scenario in Fig. 5. Compared with the baseline (first row) and tr-RoPE only (second row), our full model follows the prompts more faithfully: it introduces the new subject in 10–20s, updates the content in 20–30s to "create a small burst of flame," correctly generates the action "The tourist takes a bite" in 30–40s, and smoothly transitions the scene in 50–60s to match the prompt "disappear into the crowd." This suggests that adding AR and its anchor memory preserves richer boundary-adjacent evidence, leading to higher perceptual quality while maintaining strong dynamics and achieving the best overall results.

4.7 Analysis

To further validate the effectiveness of our AR and tr-RoPE, we conduct controlled component replacements on the LongLive baseline [41]

Replacing re-cache strategies. As reported in Tab. 6, we replace the original KV re-cache of the baseline with the Flush strategy from Infinity-RoPE [42] and our AR, while keeping other settings unchanged. Flush yields only marginal improvements over re-cache, as reflected by the CLIP score increasing from 23.58 to 23.71 and the dynamic degree rising from 26.37 to 27.60, whereas our AR achieves the best performance across all metrics with a CLIP score of 24.22, a dynamic degree of 50.73, and a quality score of 79.78. This gain is attributed

Table 6: Performance of various re-cache.

Method	CLIP Score ↑	Dynamic Degree ↑	Quality Score ↑
Baseline	23.58	26.37	78.92
Flush	23.71	27.60	79.09
AR	**24.22**	**50.73**	**79.78**

Table 7: Performance of various RoPEs.

Method	CLIP Score ↑	Dynamic Degree ↑	Quality Score ↑
Baseline	23.58	26.37	78.92
Bounded	23.58	29.09	79.38
tr-RoPE	**23.71**	**74.83**	**81.72**

to AR explicitly leveraging anchor memory that integrates sink, clip, and local caches, which jointly preserve global semantics, clip evidence, and recent context for post-switch conditioning, thereby improving prompt adherence and perceptual quality.

Replacing RoPE variants. We further compare three positional encoding variants on the same baseline: Basic RoPE with unbounded indexing, Bounded RoPE that follows the pretrained positional limit (e.g., capped at 21), and our tri-region RoPE. As shown in Tab. 7, bounded RoPE provides only a modest improvement in dynamics and quality, while tri-region RoPE delivers a substantially larger gain, especially in dynamic degree (26.37→74.83), together with improved CLIP score (23.58→23.71) and quality score (78.92→81.72). These results indicate that tri-region RoPE is critical for retaining motion priors under long-horizon streaming, and it complements AR by improving motion dynamics while AR stabilizes boundary conditioning and content fidelity.

5 Conclusion

In this work, we propose Anchor Forcing, a cache-centric framework for interactive long video generation under streaming constraints. Anchor Forcing combines an anchor-guided re-cache mechanism, which builds an anchor memory from sink, junction, and local caches to stabilize post-switch conditioning, with a tri-region RoPE that uses region-specific, range-bounded indexing to better retain pretrained motion priors over long horizons. Extensive experiments on VBench show that Anchor Forcing improves perceptual quality and motion dynamics in interactive settings and consistently outperforms prior streaming baselines, achieving state-of-the-art results.

Acknowledgments

This research was supported by The National Nature Science Foundation of China (Grant Nos: 62432014, 62402417, 62273301, 62273302), in part by "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant No. 2025C02026), in part by the Key R&D Program of Ningbo (Grant Nos: 2024Z115, 2025Z035), in part by Yongjiang Talent Introduction Programme (Grant No: 2023A-197-G).

References

1. Cai, S., Yang, C., Zhang, L., Guo, Y., Xiao, J., Yang, Z., Xu, Y., Yang, Z., Yuille, A., Guibas, L., et al.: Mixture of contexts for long video generation. arXiv preprint arXiv:2508.21058 (2025)
2. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. Advances in Neural Information Processing Systems **37**, 24081–24125 (2024)
3. Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., et al.: Skyreels-v2: Infinite-length film generative model. arXiv preprint arXiv:2504.13074 (2025)
4. Chen, X., Wang, Y., Zhang, L., Zhuang, S., Ma, X., Yu, J., Wang, Y., Lin, D., Qiao, Y., Liu, Z.: Seine: Short-to-long video diffusion model for generative transition and prediction. In: The Twelfth International Conference on Learning Representations (2023)
5. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Lol: Longer than longer, scaling video generation to hour. arXiv preprint arXiv:2601.16914 (2026)
6. Dalal, K., Koceja, D., Xu, J., Zhao, Y., Han, S., Cheung, K.C., Kautz, J., Choi, Y., Sun, Y., Wang, X.: One-minute video generation with test-time training. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17702–17711 (2025)
7. Feng, R., Zhang, H., Yang, Z., Xiao, J., Shu, Z., Liu, Z., Zheng, A., Huang, Y., Liu, Y., Zhang, H.: The matrix: Infinite-horizon world generation with real-time moving control. arXiv preprint arXiv:2412.03568 (2024)
8. Gao, J., Chen, Z., Liu, X., Feng, J., Si, C., Fu, Y., Qiao, Y., Liu, Z.: Longvie: Multimodal-guided controllable ultra-long video generation. arXiv preprint arXiv:2508.03694 (2025)
9. Gu, Y., Mao, W., Shou, M.Z.: Long-context autoregressive video modeling with next-frame prediction. arXiv preprint arXiv:2503.19325 (2025)
10. Guo, Y., Yang, C., Yang, Z., Ma, Z., Lin, Z., Yang, Z., Lin, D., Jiang, L.: Long context tuning for video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17281–17291 (2025)
11. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., Panet, P., Weissbuch, S., Kulikov, V., Bitterman, Y., Melumian, Z., Bibi, O.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
12. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation. arXiv preprint arXiv:2211.13221 (2022)
13. Henschel, R., Khachatryan, L., Poghosyan, H., Hayrapetyan, D., Tadevosyan, V., Wang, Z., Navasardyan, S., Shi, H.: Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2568–2577 (2025)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
15. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
16. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench:

Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
17. Huang, Z., Zhang, F., Xu, X., He, Y., Yu, J., Dong, Z., Ma, Q., Chanpaisit, N., Si, C., Jiang, Y., Wang, Y., Chen, X., Chen, Y.C., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench++: Comprehensive and versatile benchmark suite for video generative models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025). https://doi.org/10.1109/TPAMI.2025.3633890
18. Ji, S., Chen, X., Yang, S., Tao, X., Wan, P., Zhao, H.: Memflow: Flowing adaptive memory for consistent and efficient long video narratives. arXiv preprint arXiv:2512.14699 (2025)
19. Jin, Y., Sun, Z., Li, N., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. arXiv preprint arXiv:2410.05954 (2024)
20. Kim, J., Kang, J., Choi, J., Han, B.: Fifo-diffusion: Generating infinite videos from text without training. Advances in Neural Information Processing Systems **37**, 89834–89868 (2024)
21. Li, G., Zheng, S., Xu, S., Chen, J., Li, B., Hu, X., Zhao, L., Jiang, P.T.: Magicworld: Interactive geometry-driven video world exploration. arXiv preprint arXiv:2511.18886 (2025)
22. Li, J., Fu, X., Peng, X., Chen, W., Zheng, Y., Zhao, T., Wang, J., Chen, F., Wang, X., So, H.K.H.: Train short, inference long: Training-free horizon extension for autoregressive video generation. arXiv preprint arXiv:2602.14027 (2026)
23. Li, W., Pan, W., Luan, P.C., Gao, Y., Alahi, A.: Stable video infinity: Infinite-length video generation with error recycling. arXiv preprint arXiv:2510.09212 (2025)
24. Lu, Y., Liang, Y., Zhu, L., Yang, Y.: Freelong: Training-free long video generation with spectralblend temporal attention. Advances in Neural Information Processing Systems **37**, 131434–131455 (2024)
25. Lu, Y., Yang, Y.: Freelong++: Training-free long video generation via multi-band spectralfusion. arXiv preprint arXiv:2507.00162 (2025)
26. Mao, X., Rui, S., Ying, K., Zheng, B., Li, C., Chi, M., Zhang, K.: Packforcing: Short video training suffices for long video sampling and long context inference. arXiv preprint arXiv:2603.25730 (2026)
27. Po, R., Nitzan, Y., Zhang, R., Chen, B., Dao, T., Shechtman, E., Wetzstein, G., Huang, X.: Long-context state-space video world models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8733–8744 (2025)
28. Qiu, H., Xia, M., Zhang, Y., He, Y., Wang, X., Shan, Y., Liu, Z.: Freenoise: Tuning-free longer video diffusion via noise rescheduling. arXiv preprint arXiv:2310.15169 (2023)
29. Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., Sitzmann, V.: History-guided video diffusion. arXiv preprint arXiv:2502.06764 (2025)
30. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
31. Tan, J., Yu, H., Huang, J., Xiao, J., Zhao, F.: Freepca: Integrating consistency information across long-short frames in training-free long video generation via principal component analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27979–27988 (2025)
32. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al.: Magi-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211 (2025)

33. Tian, J., Song, C., Cheng, W., Zhang, C.: Free-lunch long video generation via layer-adaptive ood correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1973–1982 (2026)
34. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description. arXiv preprint arXiv:2210.02399 (2022)
35. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
36. Wang, W., Yang, Y.: 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), https://openreview.net/forum?id=pYNl76onJL
37. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. International Journal of Computer Vision **133**(5), 3059–3078 (2025)
38. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., Xing, S., Chen, G., Pan, J., Yu, J., Wang, Y., Wang, L., Qiao, Y.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
39. Xiang, X., Duan, Z., Zhang, G., Zhang, H., Gao, Z., Wu, J., Zhang, S., Wang, T., Fan, Q., Guo, C.: Pathwise test-time correction for autoregressive long video generation. arXiv preprint arXiv:2602.05871 (2026)
40. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. arXiv preprint arXiv:2309.17453 (2023)
41. Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., Li, M., Xie, E., Chen, Y., Lu, Y., et al.: Longlive: Real-time interactive long video generation. arXiv preprint arXiv:2509.22622 (2025)
42. Yesiltepe, H., Meral, T.H.S., Akan, A.K., Oktay, K., Yanardag, P.: Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. arXiv preprint arXiv:2511.20649 (2025)
43. Yi, J., Jang, W., Cho, P.H., Nam, J., Yoon, H., Kim, S.: Deep forcing: Training-free long video generation with deep sink and participative compression. arXiv preprint arXiv:2512.05081 (2025)
44. Yin, S., Wu, C., Yang, H., Wang, J., Wang, X., Ni, M., Yang, Z., Li, L., Liu, S., Yang, F., et al.: Nuwa-xl: Diffusion over diffusion for extremely long video generation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1309–1320 (2023)
45. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T.: Improved distribution matching distillation for fast image synthesis. In: NeurIPS (2024)
46. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: CVPR (2024)
47. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: CVPR (2025)
48. Yu, J., Bai, J., Qin, Y., Liu, Q., Wang, X., Wan, P., Zhang, D., Liu, X.: Context as memory: Scene-consistent interactive long video generation with memory retrieval. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–11 (2025)

49. Zhang, L., Cai, S., Li, M., Wetzstein, G., Agrawala, M.: Frame context packing and drift prevention in next-frame-prediction video diffusion models. arXiv preprint arXiv:2504.12626 (2025)
50. Zhao, M., He, G., Chen, Y., Zhu, H., Li, C., Zhu, J.: Riflex: A free lunch for length extrapolation in video diffusion transformers. arXiv preprint arXiv:2502.15894 (2025)