Towards Memory-Efficient Autoregressive Video Generation via Instance-Specific Parametric Absorption

Xiaomeng Fu[1,3§,*], Jia Li[2§], Yiming Hu[3‡], Yong Wang[3†,‡], Hayden Kwok-Hay So[2†], Jiao Dai[1†], Xiangxiang Chu[3], and Jizhong Han[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] The University of Hong Kong
[3] AMAP, Alibaba Group

Abstract. Autoregressive (AR) streaming models have emerged as a powerful paradigm for long video generation. However, the linearly growing Key-Value (KV) cache poses a significant bottleneck, leading to memory overload and degraded inference throughput. A common compression method is to drop redundant KV tokens, which often breaks long-range dependencies, resulting in temporal flickering and identity loss. In this paper, we propose Instance-Specific Parametric Absorption (ISPA), a novel framework that shifts the KV cache compression from discarding to distilling. The core idea is to transit a subset of layers from Full-Attention (F-Layers) to memory-efficient Local-Attention (L-Layers) by "absorbing" historical context into the model's weights. Specifically, during a brief warmup phase, ISPA monitors the output discrepancy between global and local attention. At the transition point, we solve a closed-form least-squares problem to compute an instance-specific weight modulation that compensates for the missing history. Experiments across architectures (1.3B to 14B) demonstrate that ISPA can remove up to 50% of the KV cache with near-lossless visual quality. We hope this perspective encourages future work to explore parametric memory consolidation beyond external token-level cache management for streaming generative models.

Keywords: Autoregressive Video Generation · KV Cache Compression

1 Introduction

Diffusion models have become a central paradigm in visual generation, powering diverse tasks [8, 17]. Among these tasks, video generation [2, 27, 31, 38, 39, 45] is particularly challenging, as it requires both high visual fidelity and long-range temporal coherence. Recent works extend diffusion [11, 32] and flow-matching models [7, 21, 24, 26, 30] from image-level generation to video synthesis, typically

§ Equal contribution ‡ Project lead † Corresponding author.
* This work was done during an internship at AMAP, Alibaba Group.

by generating a fixed-length clip holistically [2, 27, 31, 38, 39, 45]. While effective for short videos, this paradigm requires full spatiotemporal attention across all frames, whose quadratic memory and computation costs make long or infinite video generation prohibitively expensive.

To address this limitation, the Autoregressive (AR) streaming paradigm [3, 13–15, 25, 29, 29, 37, 44, 50] has become a popular solution. By breaking a video into sequential segments, AR models generate frames one by one, using a Key-Value (KV) cache to store historical context. This method theoretically allows for videos of any length and supports real-time generation. However, the streaming process has a major drawback: to keep the video consistent over time, the KV cache must store continually accumulated history. As generation continues, this memory bank fills the GPU memory and significantly slows down the inference speed, losing the "real-time" advantage of the system.

A common way to reduce this memory pressure is to drop redundant tokens from the KV cache. While token dropping has been effective for Large Language Models (LLMs), directly applying this idea to dynamic video signals is often unstable. The key issue is that, in long video generation, the KV cache is not merely a storage buffer, but an external token memory that carries the model's accumulated visual history. Existing pruning or eviction strategies manage this memory outside the model by deciding which tokens to keep or discard. Once important historical tokens are removed, long-range dependencies can be broken, causing the model to forget the past and leading to background flickering, structural distortion, and identity loss.

These challenges lead us to rethink how context should be compressed in a streaming setting. Instead of treating the KV cache as an ever-growing external memory to be repeatedly edited, we ask whether part of this memory can be consolidated into the model itself. The streaming nature of AR generation provides exactly such an opportunity: as a video unfolds, its instance-specific temporal patterns become observable. In some layers, the accumulated history then behaves as a stable contextual bias for future attention outputs, rather than as fully dynamic token-level information. Such history does not necessarily need to remain as explicit KV tokens; it can instead be analytically absorbed into the model's linear weights through instance-specific test-time adaptation.

Based on this insight, we propose Instance-Specific Parametric Absorption (ISPA), a novel framework that reframes KV cache compression as parametric memory consolidation. ISPA transitions selected layers from Full-Attention Layers (F-Layers) to memory-efficient Local-Attention Layers (L-Layers) during inference. In a standard F-Layer, the model must query the growing KV cache to maintain temporal consistency. In contrast, an L-Layer attends only to local frames, which is efficient but risks losing historical context. Instead of simply dropping the missing history, ISPA performs Layer Absorption: it treats the difference between full and local attention as a reconstruction problem and solves for an instance-specific weight modulation in closed form. The resulting L-Layers can emulate the effect of long-range history using only local informa-

tion, effectively internalizing part of the external KV memory into the model parameters.

A major technical challenge is performing this adaptation without interrupting the real-time generation process. To eliminate the computational overhead of monitoring the necessary attention signals (both local and full attention), we design a Decomposable Attention mechanism. By leveraging the Log-Sum-Exp (LSE) states already tracked by hardware-aware kernels (e.g., FlashAttention), we can collect both local and global attention during the standard forward pass. This enables ISPA to compute the weight updates without gradient-based optimization, thereby preserving the original inference latency of the base model.

We evaluate ISPA across several state-of-the-art architectures [15, 25, 29, 29] ranging from 1.3B to 14B parameters and various tasks (text-to-vieo and speech-to-video). ISPA successfully removes up to 50% of the KV cache with near-lossless (<1%) visual quality. Moreover, we demonstrate that ISPA is intrinsically compatible with post-training quantization. By combining parametric absorption with W8A8 quantization, we achieve an aggregate 1.86× inference speedup. Our contributions are three-fold:

- We introduce ISPA, a novel streaming framework that shifts the KV cache reduction from "pruning" to "absorption". Instead of discarding historical context, ISPA distills instance-specific context into the model's weights to maintain global coherence while operating with a minimal local cache.
- We propose a non-iterative, closed-form optimization for weight modulation. By leveraging a Decomposable Attention mechanism grounded in Online Softmax, our method collects necessary signals and updates parameters during the standard forward pass with negligible overhead.
- We conduct extensive experiments on ISPA across various AR models (from 1.3B to 14B, from text-to-video to speech-to-video). Our results demonstrate that ISPA can remove up to 50% of the KV cache with near-lossless visual quality and an aggregate 1.86× speedup when combined with quantization.

2 Preliminaries

Video Diffusion Models. Most modern video generators [2, 39] are built by extending diffusion or flow models from the image domain to the video domain. A video clip is typically represented as a 4D tensor $\mathbf{X} \in \mathbb{R}^{F \times C \times H \times W}$, where F is the number of frames. Generation originates from Gaussian noise, gradually transforming into a clean video through an iterative denoising process. In Rectified Flow, the model predicts a velocity at each step that transports the current sample towards the target data distribution.

A central challenge in video generation is maintaining temporal coherence: frames must remain consistent over time. The standard solution is to process all frames jointly utilizing spatiotemporal self-attention, allowing tokens from any frame to attend to tokens from any other frame. While highly effective, this operation is notoriously expensive; attention across F frames typically scales

as $\mathcal{O}(F^2)$ in both time and memory footprint, rendering long-video generation computationally prohibitive.

Autoregressive Video Generation. To scale to longer sequence lengths, an autoregressive (AR) strategy [14, 44, 50] generates the video in smaller, consecutive chunks. Let a long video $\mathbf{X}$ be partitioned into N chunks:

$$\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\} \tag{1}$$

where each chunk contains L frames. The AR factorization is defined as:

$$p(\mathbf{X}) = p(\mathbf{x}^{(1:N)}) = \prod_{n=1}^{N} p(\mathbf{x}^{(n)} \mid \mathbf{x}^{(<n)}) \tag{2}$$

Each chunk $\mathbf{x}^{(n)}$ is generated conditioned on the previously generated history $\mathbf{x}^{(<n)}$. By maintaining a sliding window of past context, AR models can theoretically extrapolate to videos of arbitrary length, effectively shifting the bottleneck from hardware memory limits to sequential inference latency.

Conditioning on the Past via KV Injection. A defining design choice in AR video diffusion is how the current chunk's denoiser leverages historical information. A prevalent approach is conditioning via cross-chunk self-attention: at denoising step t for chunk n, the model derives queries, keys, and values from the current noisy chunk $\mathbf{x}_t^{(n)}$. To incorporate the past, the keys and values from previously generated frames are concatenated to the current chunk's respective tensors:

$$\mathbf{Q} = \mathbf{Q}_{\text{cur}}, \quad \mathbf{K} = [\mathbf{K}_{\text{past}}; \mathbf{K}_{\text{cur}}], \quad \mathbf{V} = [\mathbf{V}_{\text{past}}; \mathbf{V}_{\text{cur}}] \tag{3}$$

Here, $(\mathbf{Q}_{\text{cur}}, \mathbf{K}_{\text{cur}}, \mathbf{V}_{\text{cur}})$ originate from the current chunk $\mathbf{x}_t^{(n)}$, while $(\mathbf{K}_{\text{past}}, \mathbf{V}_{\text{past}})$ are derived from the generated history $\mathbf{x}^{(<n)}$. Specifically, $(\mathbf{K}_{\text{past}}, \mathbf{V}_{\text{past}})$ are typically truncated to include only the KV pairs from the most recent frames to ensure memory efficiency. Intuitively, the current chunk "queries" a persistent memory bank of past frames. This KV-based injection is critical for both visual quality and generation efficiency, as it determines how temporal dependencies propagate through the autoregressive diffusion process.

3 Method

To achieve exhaustive instance-level KV cache compression without compromising real-time throughput, we propose **I**nstance-**S**pecific **P**arametric **A**bsorption (**ISPA**), a streaming compression framework. Our goal is to mine the inherent temporal redundancy within each video generation instance and compress it, while maintaining strictly negligible compute overhead to preserve the realtime property of autoregressive inference.

3.1 Overall Framework

To circumvent the memory explosion of the KV cache, ISPA shifts from conventional token pruning which irreversibly discards historical data to "Absorption"

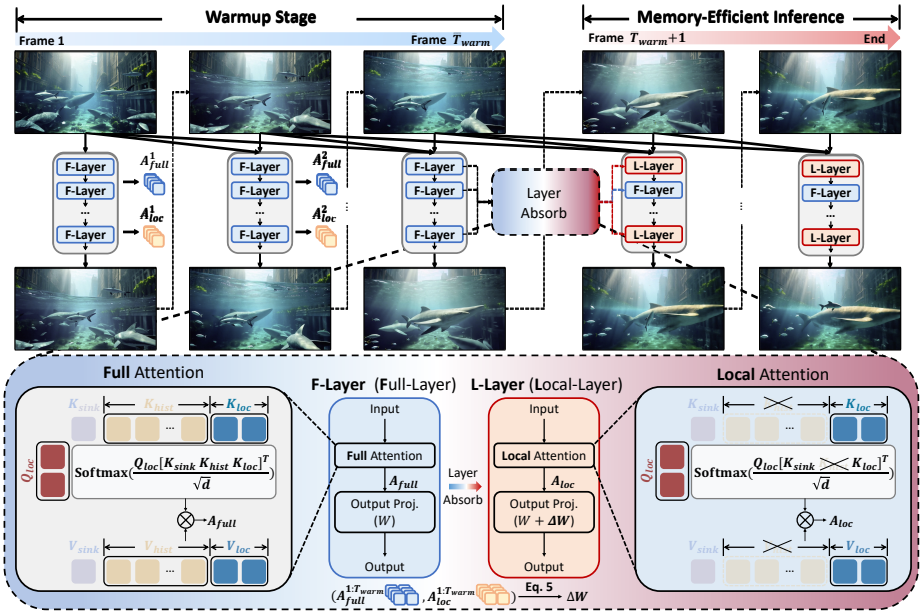

Fig. 1: The overall framework of ISPA. ISPA consists of three phases: (1) **Dual-Stream Warmup**: During the first T_{warm} frames, F-Layers collect full attention outputs A_{full} and local outputs A_{loc}. (2) **Layer Absorb**: At frame T_{warm}, we select K layers and convert them to L-Layer (Local Layer). (3) **Memory-Efficient Inference**: The selected layers permanently evict historical KV caches and utilize $W + \Delta W$ to compensate for the lost context, while the remaining $N - K$ layers stay as F-Layers.

paradigm. This paradigm distills historical context into the model's weights via *Instance-Specific* test-time adaptation, overfitting parameters to the unique temporal dynamics of the specific video sequence currently being generated. As illustrated in Figure 1, ISPA operates through a streamlined, three-stage pipeline:

1. **Dual-Stream Warmup:** During the generation of the initial T_{warm} frames, the model operates in a dual-stream configuration using **F-Layers** (Full-Layers). We simultaneously collect the "Full" attention activation $\mathbf{A}_{full}$ (utilizing the whole historical context) and the "Local" activation $\mathbf{A}_{loc}$ (utilizing only current and sink tokens) via an optimized kernel-level decomposition.
2. **Layer Absorb:** At the dividing frame T_{warm}, the accumulated dual-stream activations serve as the basis for weight adaption. We formulate the absorption as a reconstruction problem, computing the optimal weight modulation ΔW for the output projection to bridge the gap between local and full attention. Rather than treating all layers equally, we employ a dynamic ranking mechanism to identify K "absorbable" layers that most accurately reconstruct global dependencies through weight modulation.
3. **Memory-Efficient Inference:** For all subsequent frames $t > T_{warm}$, the designated K layers transition into **L-Layers** (Local-Layers). These layers

completely discard their historical KV cache $(\mathcal{K}_{hist}, \mathcal{V}_{hist})$, retaining only the local context. The absence of this history is effectively compensated for by the newly updated weights $W_{new} = W + \Delta W$. The remaining $N - K$ layers continue as F-Layers to maintain global coherence.

Although we present this warmup-to-absorption transition as a one-shot pipeline for clarity, the same procedure can be re-triggered when the video enters a new temporal regime, allowing ISPA to recalibrate the absorbed layers and their weight modulation for the new context.

3.2 Layer Absorb via Full Attention Reconstruction

The computational core of ISPA lies in absorbing the historical context of a layer into its weights. We formulate this as a unified least-squares optimization problem, solved at the end of the warmup phase.

As shown in the "F-Layer" block of Figure 1, for each diffusion transformer layer $l \in \{1, 2, ..., N\}$, we monitor the internal activations across T_{warm} frames. Let $\mathbf{A}_{loc,l}^{1:T_{warm}} \in \mathbb{R}^{T_{warm} \times D}$ denote the attention output immediately preceding the final linear projection layer W, computed utilizing *only* the local context (current frame and sink frames). Let $\mathbf{A}_{full,l}^{1:T_{warm}} \in \mathbb{R}^{T_{warm} \times D}$ represent the ground-truth "full" output of that same attention block. **For notational brevity, we omit the layer index and the temporal superscripts in the following derivation.** Our objective is to discover a weight modulation matrix ΔW such that the local-stream output after projection matches the full-stream output. Specifically, we target the linear projection layer for adaption as it is the immediate downstream component of the attention block. By compensating for the context loss at this point of origin, we minimize the propagation of reconstruction errors to subsequent layers. The optimization is formulated as:

$$\min_{\Delta W} ||\mathbf{A}_{loc}(W + \Delta W) - \mathbf{A}_{full} W||_2^2 + \lambda ||\Delta W||_2^2 \tag{4}$$

By defining the target residual signal as $\mathbf{R} = (\mathbf{A}_{full} - \mathbf{A}_{loc})W$, the optimal ΔW that absorbs the truncated history can be derived via the closed-form regression solution:

$$\Delta W = ((\mathbf{A}_{loc})^\top \mathbf{A}_{loc} + \lambda \mathbf{I})^{-1} (\mathbf{A}_{loc})^\top \mathbf{R} \tag{5}$$

The reconstruction error (feasibility score) for each layer is then defined as the Frobenius norm of the residual:

$$\epsilon = ||\mathbf{A}_{loc}(W + \Delta W) - \mathbf{A}_{full} W||_F^2 \tag{6}$$

A major advantage of this formulation is its strictly **non-iterative** nature. Unlike traditional knowledge distillation that necessitates backpropagating gradients through massive computational graphs, ISPA only requires summing the outer products of activations ($\mathbf{A}_{loc}^\top \mathbf{A}_{loc}$ and $\mathbf{A}_{loc}^\top \mathbf{R}$) during the standard forward pass. This reduces optimization to a single $D \times D$ matrix inversion precisely at frame T_{warm}. Consequently, the adaptation signal is harvested and applied without ever initializing a gradient-based optimizer, rigorously preserving the real-time inference pipeline.

3.3 Dynamic Layer Selection

Having computed the candidate update ΔW and error metric ϵ for all layers at the end of T_{warm}, we dynamically route which layers will transition to a KV cache-free regime.

Rather than enforcing a fixed subset of layers across all prompts, ISPA executes Instance-Specific Selection. Given a hardware budget of K layers designated for KV cache eviction, we rank layers in ascending order based on ϵ. The K layers exhibiting the lowest ϵ are designated as **L-Layers**. For these layers, we inject the update $W_{new} = W + \Delta W$ and **permanently** evict their historical KV cache $\mathcal{K}_{hist}$. The remaining $N - K$ layers, which evidently handle complex temporal dependencies that resist linear parametric absorption, remain as **F-Layers** to guarantee generation quality.

This dynamic routing is essential because diverse video sequences exhibit significantly different temporal dynamics and these dynamics heterogeneously distribute across the model's hierarchy. By delaying the architectural decision until T_{warm}, ISPA autonomously tailors the model's topology to the specific instance, a phenomenon we term test-time structural elasticity. Once the topology is locked, the model's parameters are updated in-place. For all subsequent frames $t > T_{warm}$, L-Layers execute using only localized KV pairs, yet their modified W_{new} simulates the continuous presence of the evicted history.

Importantly, this calibration is not permanently tied to the initial scene. The one-shot schedule described above is the default setting used in our experiments, but the same procedure can be re-triggered when the video undergoes a large scene transition or temporal distribution shift. In such cases, ISPA briefly re-enters the dual-stream warmup stage, recomputes the layer-wise reconstruction errors, and refreshes both the absorbed layer set and the corresponding ΔW for the new context. Since this adaptation is closed-form and lightweight, event-triggered recalibration preserves the streaming nature of inference while avoiding stale weight modulation under changing scenes.

3.4 Real-time Efficiency Optimization via Decomposable Attention

A potential bottleneck of ISPA is the necessity to simultaneously collect two distinct attention signals (the localized attention $\mathbf{A}_{loc}$ for ΔW computation and the full attention $\mathbf{A}_{full}$ for standard inference) during the warmup phase. A naive implementation would dual-process the attention mechanism, degrading real-time performance. To eliminate this overhead, we introduce a *Decomposable Attention* mechanism grounded in the Online Softmax theorem.

For a video sequence at step $t \in [1, T_{warm}]$, we define the **Local Context** as the union of the initial frame (serving as the "sink frame", $K_{sink} = K_1$) and the current frame's KV ($K_{loc} = K_t$). Formally, the key and value sets are partitioned as follows:

$$\mathcal{K}_{loc} = \{K_1, K_t\}, \quad \mathcal{V}_{loc} = \{V_1, V_t\} \tag{7}$$

$$\mathcal{K}_{hist} = \{K_2, \ldots, K_{t-1}\}, \quad \mathcal{V}_{hist} = \{V_2, \ldots, V_{t-1}\} \tag{8}$$

The exhaustive KV cache is the disjoint union: $\mathcal{K}_{full} = \mathcal{K}_{loc\_total} \cup \mathcal{K}_{hist}$.

By exploiting the Log-Sum-Exp (LSE) state inherently tracked by modern hardware-aware kernels [5, 6, 52, 53] (e.g., FlashAttention), we decouple the attention into two independent branches and fuse them post-hoc. For each layer l, we compute the local and historical outputs along with their corresponding LSE values:

$$(\mathbf{O}_{loc}, LSE_{loc}) = \text{FlashAttn}(\mathbf{Q}_t, \mathcal{K}_{loc}, \mathcal{V}_{loc}) \tag{9}$$

$$(\mathbf{O}_{hist}, LSE_{hist}) = \text{FlashAttn}(\mathbf{Q}_t, \mathcal{K}_{hist}, \mathcal{V}_{hist}) \tag{10}$$

$\mathbf{O}_{loc}$ is mathematically identical to $\mathbf{A}_{loc}$. To perfectly reconstruct the Full output $\mathbf{A}_{full}$ without recomputing attention, we apply a weighted fusion driven by the LSE values:

$$LSE_{full} = \log\left(\exp(LSE_{loc}) + \exp(LSE_{hist})\right) \tag{11}$$

$$\mathbf{A}_{full} = \exp(LSE_{loc} - LSE_{full}) \cdot \mathbf{O}_{loc} + \exp(LSE_{hist} - LSE_{full}) \cdot \mathbf{O}_{hist} \tag{12}$$

This dual-path decomposition ensures that the warmup phase is nearly "free" in terms of FLOPs. The total dot-product operations remain almost identical to a standard attention pass as the query vector is simply partitioned across the same total number of KV pairs. Since the fusion involves only scalar operations in SRAM, we avoid materializing massive intermediate attention matrices, preserving the memory efficiency of FlashAttention. The final fusion accounts for less than 1% of execution time, ensuring a smooth transition from full-cache inference to parametric-absorbed inference.

4 Experiments

4.1 Experimental Settings

Models and Benchmarks. To evaluate the generalizability and robustness of our proposed ISPA, we conduct extensive experiments across four representative autoregressive video generation models: **LongLive** [44], **Reward-Forcing** [25] (denoted as **Reward**), **Krea-Realtime** [29] (denoted as **Krea**), and **LiveAvatar** [15]. These models cover a diverse range of model size (from 1.3B to 14B) and generation task (including text-to-video and speech-to-video). The selection of these heterogeneous architectures, spanning from general-purpose synthesis to specialized human-centric generation, allows us to validate ISPA's performance across varying modeling paradigms. For long video evaluation, we primarily employ the **VBench-Long** benchmark [16,55]. In addition to the standard prompts provided by VBench, we incorporate complex full-length prompts from **MovieGen** [31] to further challenge the models' capability in maintaining long-range temporal consistency and semantic alignment. This combination ensures a comprehensive assessment that covers both basic motion quality and high-level narrative coherence. All ablation studies and analytical experiments are conducted based on the LongLive model as default backbone.

Table 1: Quantitative evaluation on the VBench **5s** benchmarks (prompts from MovieGen). **Aes.**: aesthetic quality; **Back.**: background consistency; **Dyn.**: dynamic degree; **Image.**: imaging quality; **Mot.**: motion smoothness; **Sub.**: subject consistency; **Temp.**: temporal flickering. **ISPA-x** indicates that a fraction x of F-Layers are replaced by L-Layers (i.e., $K = xN$).

Model	Method	KVCache	VBench Metrics						
			Aes.↑	Back.↑	Dyn.↑	Image.↑	Mot.↑	Sub.↑	Temp.↑
LongLive (1.3B)	Vanilla	100%	64.43	95.76	40.77	70.74	98.92	95.55	97.87
	ISPA-0.3	72.5%	64.48(+0.05)	95.80(+0.04)	40.60(-0.17)	70.64(-0.10)	98.91(-0.01)	95.53(-0.02)	97.87(+0.00)
	ISPA-0.4	63.3%	64.47(+0.04)	95.80(+0.04)	40.60(-0.17)	70.64(-0.10)	98.91(-0.01)	95.53(-0.02)	97.87(+0.00)
	ISPA-0.5	54.2%	64.48(+0.05)	95.78(+0.02)	40.60(-0.17)	70.70(-0.04)	98.92(+0.00)	95.59(+0.04)	97.86(-0.01)
Reward (1.3B)	Vanilla	100%	62.82	94.73	68.96	68.65	98.08	93.76	96.34
	ISPA-0.3	73.3%	62.83(+0.01)	94.74(+0.01)	69.50(+0.54)	68.65(+0.00)	98.08(+0.00)	93.76(+0.00)	96.34(+0.00)
	ISPA-0.4	64.4%	62.82(+0.00)	94.74(+0.01)	69.17(+0.21)	68.65(+0.00)	98.08(+0.00)	93.76(+0.00)	96.34(+0.00)
	ISPA-0.5	55.6%	62.85(+0.03)	94.73(+0.00)	69.60(+0.64)	68.65(+0.00)	98.08(+0.00)	93.76(+0.00)	96.34(+0.00)
Krea (14B)	Vanilla	100%	64.50	95.40	77.12	70.86	98.62	94.90	96.46
	ISPA-0.3	75.0%	64.51(+0.01)	95.16(-0.24)	76.65(-0.47)	71.03(+0.17)	98.71(+0.09)	94.66(-0.24)	96.91(+0.45)
	ISPA-0.4	66.7%	64.66(+0.16)	94.68(-0.72)	76.15(-0.97)	70.70(-0.16)	98.80(+0.18)	94.02(-0.88)	97.14(+0.68)
	ISPA-0.5	58.3%	64.93(+0.43)	94.06(-1.34)	76.23(-0.89)	71.01(+0.15)	98.72(+0.10)	92.32(-2.58)	97.04(+0.58)

Table 2: Quantitative evaluation on the VBench **30s** benchmarks (prompts from MovieGen). Notations and metric definitions follow Tab 1.

Model	Method	KVCache	VBench Metrics						
			Aes.↑	Back.↑	Dyn.↑	Image.↑	Mot.↑	Sub.↑	Temp.↑
LongLive (1.3B)	Vanilla	100%	63.59	96.65	41.93	70.15	98.91	96.10	97.92
	ISPA-0.3	72.5%	63.32(-0.27)	96.48(-0.17)	41.38(-0.55)	69.38(-0.77)	98.90(-0.01)	95.77(-0.33)	97.95(+0.03)
	ISPA-0.4	63.3%	63.05(-0.54)	96.29(-0.36)	42.37(+0.44)	69.05(-1.10)	98.88(-0.03)	95.52(-0.58)	97.90(-0.02)
	ISPA-0.5	54.2%	62.98(-0.61)	96.28(-0.37)	41.38(-0.55)	69.05(-1.10)	98.88(-0.03)	95.61(-0.49)	97.89(-0.03)
Reward (1.3B)	Vanilla	100%	60.82	94.97	66.92	66.79	98.05	93.43	96.49
	ISPA-0.3	73.3%	60.82(+0.00)	94.98(+0.01)	67.24(+0.32)	66.79(+0.00)	98.05(+0.00)	93.43(+0.00)	96.49(+0.00)
	ISPA-0.4	64.4%	60.72(-0.10)	95.01(+0.04)	65.45(-1.47)	66.68(-0.11)	98.09(+0.04)	93.51(+0.08)	96.58(+0.09)
	ISPA-0.5	55.6%	60.45(-0.37)	94.95(-0.02)	65.09(-1.83)	66.55(-0.24)	98.10(+0.05)	93.36(-0.07)	96.64(+0.15)
Krea (14B)	Vanilla	100%	62.15	94.64	75.13	71.20	98.84	94.85	96.85
	ISPA-0.3	75.0%	62.54(+0.39)	95.32(+0.68)	74.47(-0.66)	70.46(-0.74)	98.66(-0.18)	94.91(+0.06)	96.91(+0.06)
	ISPA-0.4	66.7%	63.29(+1.14)	95.17(+0.53)	74.80(-0.33)	70.27(-0.93)	98.81(-0.03)	94.76(-0.09)	97.26(+0.41)
	ISPA-0.5	58.3%	63.93(+1.78)	94.63(-0.01)	75.27(+0.14)	70.12(-1.08)	98.73(-0.11)	93.81(-1.04)	97.12(+0.27)

Implementation Details. During the first 12 latent frames ($T_{warm} = 12$), the model generates video using standard KV-Cache while simultaneously collecting $\mathbf{A}_{loc}$ and $\mathbf{A}_{full}$. At the switching point T_{warm}, we apply a **closed-form update** to the model parameters to absorb the historical context, after which the explicit KV-Cache is pruned. For the closed-form adaptation (Eq. 5), we employ

a regularization parameter $\lambda = 0.001$ by default to balance the preservation of pre-trained knowledge and the integration of new temporal context.

4.2 Main Results

Quantitative Results. Tab 1 and 2 evaluates ISPA's ability to internalize historical context without suffering from "temporal amnesia". On standard 5s videos (1.3B models), ISPA achieves near lossless compression: evicting almost 50% KV caches causes negligible deviations. Crucially, when scaling to 30s long-form generation, ISPA not only maintains minor degradation on 1.3B models but actively *improves* Aesthetic Score and Temporal Consistency on the 14B Krea-Realtime model. We attribute this long-video improvement to our analytic absorption ($W_{new} = W + \Delta W$). Explicit attention over massive historical tokens often introduces high-frequency noise, corrupting diffusion trajectories. By forcing half of the layers to summarize context into static weights, ISPA explicitly filters this noise. While slightly blurring local token matching (minor Subject drop), it powerfully reinforces global semantic and temporal stability, acting as a macroscopic regularizer rather than just an efficiency optimization.
Qualitative Results. Fig. 2 visually validates ISPA's robustness in 30-second long-form generation. Across 1.3B to 14B models, ISPA aligns closely with full-cache baselines. Notably, at the 30-second mark (rightmost frames), ISPA consistently preserves intricate details. This confirms that our analytic absorption successfully distills historical context into projection weights, achieving substantial memory compression without visual degradation. Furthermore, the quantitative comparisons on the right panel highlight ISPA's superior memory efficiency. Across various architectures, ISPA consistently reduces GPU memory. Specifically, in the most Krea model (14B), ISPA achieves a remarkable reduction of 23.7GB in peak memory. This significant margin demonstrates that ISPA not only maintains high fidelity synthesis but also makes long-video generation feasible on consumer-grade hardware by effectively capping memory growth.

4.3 Ablation Study

The impact of Layer Absorb and Sink Frame. To analyze the mechanism of ISPA, we conduct an ablation study across two key dimensions: weight adaptation and attention anchoring (Fig. 3). In the "No Absorb" setting (middle row), we evaluate the performance of simple local attention without our analytic ΔW. This leads to a catastrophic "Subject Mismatch," where the creature's morphology drifts significantly as the generation progresses. This highlights that our instance-specific adaptation is not merely an auxiliary update, but a necessary operation to reconstruct the missing global receptive field. We also evaluate the performance with parametric adaptation while removing the sink frame ("No Sink", bottom row), which causes the model to generate "Over-Smoothed" results. We attribute this to the fact that the sink frame provides a stable basis for softmax normalization; its absence forces the attention mechanism to redistribute scores across less relevant local tokens, resulting in blurry, low-fidelity

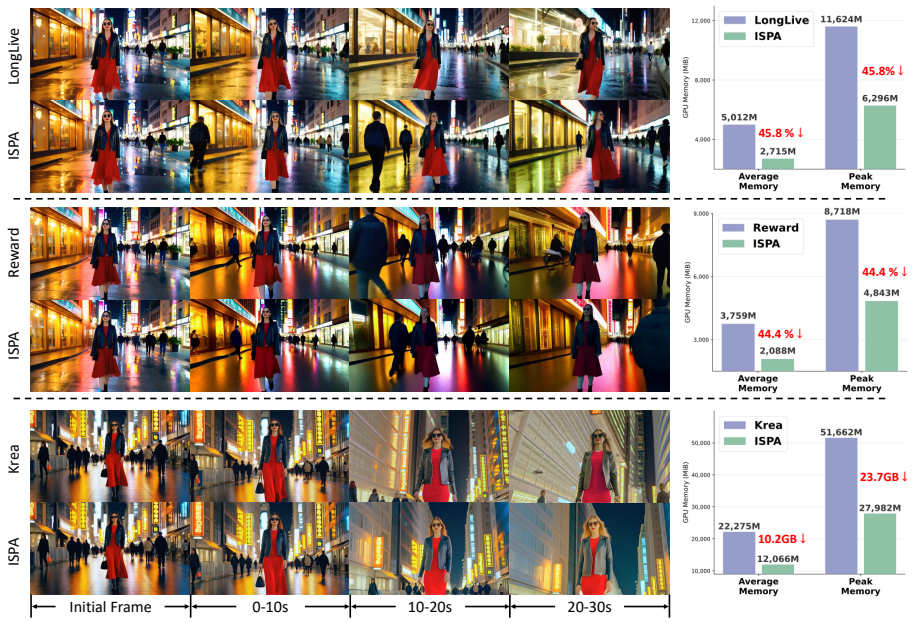

Fig. 2: Left: Qualitative comparison of 30-second long-form video generation. Right: GPU memory usage comparison.

outputs. The combination between parametric absorption (for identity) and sink tokens (for stability) allows ISPA to match the quality of full-cache inference while maintaining a constant memory footprint.

The impact of the number of L-Layers K. Figure 4 evaluates the robustness of ISPA across varying compression ratios (defined by the number of L-Layers K). For $K \leq 0.5N$, ISPA maintains a performance nearly identical to the uncompressed baseline. Notably, in late-stage generation (>15s), ISPA with moderate K even slightly outperforms the baseline in Aesthetic Quality. However, the video quality decreases as K increase further (when $K = 0.7N$). This reveals a efficiency-utility tradeoff: although linear absorption can preserve global context, a sufficient number of F-Layers is essential for modeling nonlinear, dynamic interactions. ISPA's dynamic selection mechanism effectively captures the tradeoff by preserving the most nonlinear layers, ensuring stability when the budge K remains within a safe limit ($K \leq 0.5N$).

The impact of warmup frame T_{warm}. Figure 5 investigates the sensitivity of our closed-form adaptation to the warmup frame. At $T_{warm} = 6$, performance drops sharply across all metrics. We attribute this to **information starvation**: with insufficient temporal diversity, the derived ΔW fits to the narrow distribution of the first several frames and fails to generalize as the video's semantics evolve in later stages. Conversely, performance saturates and closely tracks the uncompressed baseline once $T_{warm} \geq 12$, confirming that a brief temporal window provides sufficient statistical evidence to capture the core temporal dy-

Fig. 3: Ablation study on core components. Top: Our full method maintains identity consistency and visual sharpness. Middle: Removing parametric adaptation (No Absorb) leads to subject morphing over time. Bottom: Omitting initial KV pairs (No Sink) causes oversmoothness.

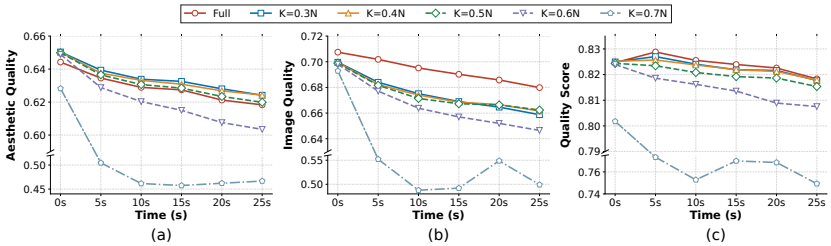

Fig. 4: Ablation on KV cache eviction ratios over time. We evaluate (a) Aesthetic Quality, (b) Image Quality and Overall (c) Quality Score across a 30-second generation horizon.

namics. While $T_{warm} = 24$ remains stable, it delays the benefits of KV cache compression. Consequently, we adopt $T_{warm} = 12$ as the optimal trade-off between adaptation accuracy and immediate memory efficiency.

Impact of Warm-up Duration T_{warmp}. Figure 5 evaluates the sensitivity of our closed-form adaptation to T_{warmp}. At T=6, performance degrades rapidly. We attribute this to *information starvation*: lacking temporal diversity, the derived ΔW overfits to initial frames and fails to generalize to later semantic shifts. Conversely, for T$\geq$12, performance stabilizes and closely tracks the Full baseline, confirming that a brief window (2-3s) suffices to capture core temporal dynamics. Furthermore, $T \in [12, 18]$ often outperforms the baseline in late stages, reinforcing that properly calibrated absorption filters long-range attention noise. Finally, while $T = 24$ is stable, it delays KV cache eviction. We therefore adopt $T = 12$ as the optimal trade-off, balancing mathematical convergence with immediate memory savings.

4.4 Further Analysis

Extension to Speech-to-Video (S2V) model. To further validate the versatility of ISPA beyond text-to-video generation, we integrate it into the speech-

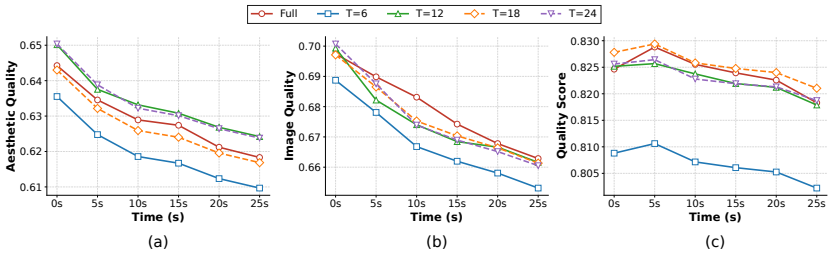

Fig. 5: Impact of warm-up duration (T_{warm}). We compare parametric adaptation using different initial frame counts. $T_{warm} \geq 12$ effectively prevents information starvation, matching or exceeding the full-cache baseline's long-term stability.

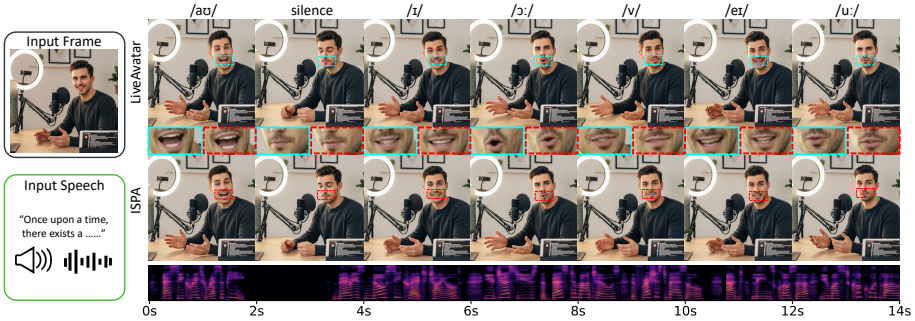

Fig. 6: Generalization to Speech-to-Video (S2V) model. We integrate ISPA into the LiveAvatar for streaming talking-head synthesis. Given a single reference image and a continuous speech input, ISPA (bottom row) maintains nearly identical visual quality and viseme accuracy compared to the full-cache baseline (top row).

to-video model (LiveAvatar). Speech-to-video synthesis is a demanding streaming task where the model must maintain strict identity consistency and precise lip-sync over hundreds of frames. As illustrated in Figure 6, we compare ISPA against the full-cache baseline (LiveAvatar). It can be observed that ISPA successfully preserves the speaker's identity and fine-grained facial expressions. Crucially, the mouth shapes (visemes) remain accurately synchronized with the input phonemes (e.g., /v/, /u:/). This indicates that our parametric update ΔW effectively "remembers" the specific facial dynamics of the speaker, confirming ISPA as a highly adaptable solution for diverse autoregressive architectures.

Compatibility with Post-Training Quantization (PTQ). While ISPA effectively mitigates the KV cache bottleneck in the attention mechanism, the computational load of linear projection layers remains a significant factor for end-to-end throughput. To address this, we explore the compatibility of ISPA with 8-bit weight and activation quantization (W8A8), aiming for a holistic acceleration of the Diffusion Transformer backbone. As illustrated in Fig. 7 (b), the standard ISPA delivers a 1.35× overall speedup by drastically optimizing the

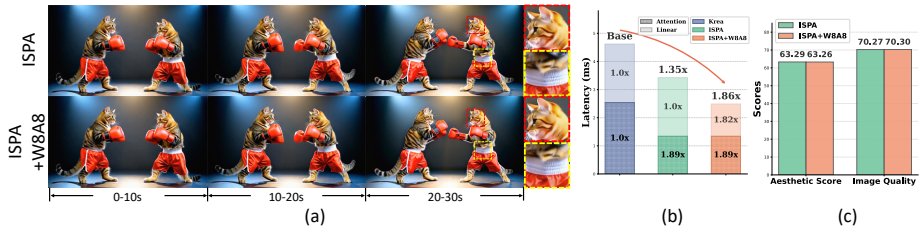

Fig. 7: Compatibility with Post-Training Quantization (PTQ). (a). Visual comparison between the standard ISPA and its quantized version (ISPA+W8A8). (b). Overall latency breakdown. (c). Quantitative stability analysis based on generation quality.

attention blocks (achieving 1.89× within the operator). By further integrating W8A8 quantization for the linear layers, we reach an aggregate speedup of **1.86**× compared to the full-precision baseline. Importantly, our parametric update ΔW is derived through a stable, closed-form reconstruction, which inherently maintains the activation distributions. This property makes ISPA highly quantization-friendly. As shown in the qualitative results (Fig. 7(a)), ISPA+W8A8 exhibits no visual degradation or structural drifting. The quantitative scores for aesthetics and image quality (Fig. 7(c)) remain virtually unchanged after quantization. These results suggest that ISPA can serve as a robust foundation for ultra-efficient streaming systems.

5 Related Works

Diffusion and Flow-Matching for Video Generation. Recent advancements in video generation are primarily driven by scaling Diffusion Models and Rectified Flow to the temporal dimension. Building upon the success of Diffusion Transformers (DiT) [30], state-of-the-art models such as Sora, Latte [27], and MovieGen [31] have demonstrated remarkable capabilities in synthesizing high-fidelity videos. These frameworks typically employ 3D spatiotemporal attention to capture inter-frame dependencies. However, since the self-attention complexity grows quadratically with the number of tokens $\mathcal{O}(F^2)$, these holistic, offline generation paradigms are fundamentally limited by hardware memory, making the synthesis of long-form or high-resolution videos computationally prohibitive.
Autoregressive (AR) Video Generation. To circumvent the memory bottleneck of offline models, the Autoregressive (AR) paradigm [3, 18, 23, 36] has emerged as the standard for long-video synthesis. By partitioning a video into sequential chunks and generating them conditioned on previous frames, AR models [4, 10, 19, 28, 41, 42, 46, 47, 56] theoretically enable infinite video extrapolation. A pioneer work is Causvid [50], which distills non-causal models into a autoregressive few-step model with DMD [48,49] and trajectory distillation [33,34]. The temporal condition is typically implemented via Key-Value (KV) cache injection, where the current chunk attends to a persistent memory bank of historical tokens. While effective, this strategy shifts the burden to memory capacity; as the

video progresses, the linear accumulation of the KV cache eventually leads to memory saturation and significant inference latency spikes.

KV Cache Compression in LLM. KV cache compression has been extensively studied in Large Language Models (LLMs), where the main goal is to reduce the memory cost of long-context decoding while preserving language modeling quality. A common line of work [1, 12, 20, 22, 35, 40, 43] exploits the sparsity structure of attention. StreamingLLM [43] and related studies observe that a small number of initial tokens, known as attention sinks [9,43,51], receive consistently high attention and are important for stabilizing streaming inference; therefore, they keep these sink tokens together with recent tokens while evicting the middle context. Other methods estimate token importance from attention statistics. For example, H2O [54] retains heavy-hitter tokens that accumulate large attention scores, while SnapKV [20] uses an observation window to identify task-relevant KV tokens before generation. More recent methods further refine this idea with query-aware or multi-stage compression, such as QUEST [35] and RocketKV [1], which adapt KV selection according to the current decoding query or combine coarse and fine pruning stages.

Despite their effectiveness in LLMs, these methods still manage KV cache as an external token memory by deciding which cached states to preserve or discard. This paradigm is fragile for autoregressive video generation, where historical KV tokens encode dense visual states and long-range temporal consistency. ISPA therefore departs from token-level cache management by absorbing part of the historical context into instance-specific weight modulation.

6 Conclusion

In this work, we rethink the challenge of long-range dependency in video generation and propose ISPA. Rather than viewing KV cache compression merely as a problem of selecting or discarding tokens, we treat the growing KV cache as an external memory of the streaming model and explore how part of this memory can be internalized into the model itself. Through instance-specific parametric absorption, ISPA consolidates historical context into lightweight weight modulation at test time, allowing selected layers to operate with local attention while preserving long-range temporal coherence.

Enabled by our Decomposable Attention mechanism and closed-form adaptation, this process provides a persistent, constant-memory representation of video history without gradient-based optimization. Extensive evaluations across DiT-based architectures ranging from 1.3B to 14B parameters demonstrate that ISPA removes up to 50% of the KV cache with near-lossless visual fidelity, while its compatibility with quantization delivers an aggregate 1.86× inference speedup. Beyond these empirical gains, we hope this work encourages the community to look beyond external token-level cache management and further explore parametric memory consolidation as a general route toward efficient streaming generative models.

References

1. Behnam, P., Fu, Y., Zhao, R., Tsai, P.A., Yu, Z., Tumanov, A.: Rocketkv: Accelerating long-context llm inference via two-stage kv cache compression. In: International Conference on Machine Learning. pp. 3358–3392. PMLR (2025)
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
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. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Self-forcing++: Towards minute-scale high-quality video generation. arXiv preprint arXiv:2510.02283 (2025)
5. Dao, T.: Flashattention-2: Faster attention with better parallelism and work partitioning. In: The Twelfth International Conference on Learning Representations
6. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in neural information processing systems **35**, 16344–16359 (2022)
7. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
8. Fu, X., Li, J.: Tcfg: Truncated classifier-free guidance for efficient and scalable text-to-image acceleration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18552–18562 (2025)
9. Gu, X., Pang, T., Du, C., Liu, Q., Zhang, F., Du, C., Wang, Y., Lin, M.: When attention sink emerges in language models: An empirical view. In: The Thirteenth International Conference on Learning Representations
10. Guo, H., Jia, Z., Li, J., Li, B., Cai, Y., Wang, J., Li, Y., Lu, Y.: Efficient autoregressive video diffusion with dummy head. arXiv preprint arXiv:2601.20499 (2026)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
12. Hooper, C.R.C., Kim, S., Mohammadzadeh, H., Maheswaran, M., Zhao, S., Paik, J., Mahoney, M.W., Keutzer, K., Gholami, A.: Squeezed attention: Accelerating long context length llm inference. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 32631–32652 (2025)
13. Huang, J., Ye, Z., Hu, X., He, T., Zhang, G., Shi, S., Bian, J., Jiang, L.: Live: Long-horizon interactive video world modeling. arXiv preprint arXiv:2602.03747 (2026)
14. 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)
15. Huang, Y., Guo, H., Wu, F., Zhang, S., Huang, S., Gan, Q., Liu, L., Zhao, S., Chen, E., Liu, J., et al.: Live avatar: Streaming real-time audio-driven avatar generation with infinite length. arXiv preprint arXiv:2512.04677 (2025)

16. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
17. Li, J., Fu, X., Gao, Y., Wang, J., Wang, X., So, H.K.H.: Rethinking conditioning in diffusion models: Dynamic token scheduling for efficient and aligned text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4160–4169 (2026)
18. 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)
19. 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)
20. Li, Y., Huang, Y., Yang, B., Venkitesh, B., Locatelli, A., Ye, H., Cai, T., Lewis, P., Chen, D.: Snapkv: Llm knows what you are looking for before generation. Advances in Neural Information Processing Systems **37**, 22947–22970 (2024)
21. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations
22. Liu, A., Liu, J., Pan, Z., He, Y., Haffari, G., Zhuang, B.: Minicache: Kv cache compression in depth dimension for large language models. Advances in Neural Information Processing Systems **37**, 139997–140031 (2024)
23. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time. arXiv preprint arXiv:2509.25161 (2025)
24. Liu, X., Gong, C., et al.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations
25. Lu, Y., Zeng, Y., Li, H., Ouyang, H., Wang, Q., Cheng, K.L., Zhu, J., Cao, H., Zhang, Z., Zhu, X., et al.: Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. arXiv preprint arXiv:2512.04678 (2025)
26. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: European Conference on Computer Vision. pp. 23–40. Springer (2024)
27. Ma, X., Wang, Y., Chen, X., Jia, G., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024)
28. Ma, Y., Zheng, X., Xu, J., Xu, X., Ling, F., Zheng, X., Kuang, H., Li, H., Wang, X., Xiao, X., et al.: Flow caching for autoregressive video generation. arXiv preprint arXiv:2602.10825 (2026)
29. Millon, E.: Krea realtime 14b: Real-time video generation (2025), `https://github.com/krea-ai/realtime-video`
30. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
31. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)

32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
33. Song, Y., Dhariwal, P.: Improved techniques for training consistency models. In: The Twelfth International Conference on Learning Representations
34. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: International Conference on Machine Learning. pp. 32211–32252. PMLR (2023)
35. Tang, J., Zhao, Y., Zhu, K., Xiao, G., Kasikci, B., Han, S.: Quest: Query-aware sparsity for efficient long-context llm inference. In: Forty-first International Conference on Machine Learning
36. Team, R., Gao, Z., Wang, Q., Zeng, Y., Zhu, J., Cheng, K.L., Li, Y., Wang, H., Xu, Y., Ma, S., et al.: Advancing open-source world models. arXiv preprint arXiv:2601.20540 (2026)
37. 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)
38. 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 descriptions. In: International Conference on Learning Representations
39. 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)
40. Wang, A., Chen, H., Tan, J., Zhang, K., Cai, X., Lin, Z., Han, J., Ding, G.: Prefixkv: Adaptive prefix kv cache is what vision instruction-following models need for efficient generation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
41. Wu, H., Wu, D., He, T., Guo, J., Ye, Y., Duan, Y., Bian, J.: Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. arXiv preprint arXiv:2507.07982 (2025)
42. Wu, X., Zhang, G., Xu, Z., Zhou, Y., Lu, Q., He, X.: Pack and force your memory: Long-form and consistent video generation. arXiv preprint arXiv:2510.01784 (2025)
43. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. arXiv preprint arXiv:2309.17453 (2023)
44. 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)
45. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
46. 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)
47. 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)
48. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, B.: Improved distribution matching distillation for fast image synthesis. Advances in neural information processing systems **37**, 47455–47487 (2024)

49. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024)
50. 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: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22963–22974 (2025)
51. Yu, Z., Wang, Z., Fu, Y., Shi, H., Shaikh, K., Lin, Y.C.: Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. In: International Conference on Machine Learning. pp. 57659–57677. PMLR (2024)
52. Zhang, J., Huang, H., Zhang, P., Wei, J., Zhu, J., Chen, J.: Sageattention2: Efficient attention with thorough outlier smoothing and per-thread int4 quantization. In: International Conference on Machine Learning. pp. 75097–75119. PMLR (2025)
53. Zhang, J., Zhang, P., Zhu, J., Chen, J., et al.: Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. In: The Thirteenth International Conference on Learning Representations
54. Zhang, Z., Sheng, Y., Zhou, T., Chen, T., Zheng, L., Cai, R., Song, Z., Tian, Y., Ré, C., Barrett, C., et al.: H2o: Heavy-hitter oracle for efficient generative inference of large language models. Advances in Neural Information Processing Systems **36**, 34661–34710 (2023)
55. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Gu, L., Zhang, Y., He, J., Zheng, W.S., et al.: Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755 (2025)
56. Zhu, H., Zhao, M., He, G., Su, H., Li, C., Zhu, J.: Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. arXiv preprint arXiv:2602.02214 (2026)