

SyncCache: Exploiting Asymmetric Dynamics for Fast Audio-Driven Portrait Animation

Juncheng Ma¹, Yuxuan Du¹, Yanan SUN², Zhening Xing²,
Changlin Li³, Zhenyu Tang¹, Bo Li⁴, Peng-Tao Jiang⁴, Li Yuan¹,
Daquan Zhou^{✉1}, and Yonghong Tian^{✉1}

¹ Shenzhen Graduate School, Peking University, China,
junchengma25@stu.pku.edu.cn
² Shanghai AI Laboratory, China,
³ Tencent Hunyuan, China,
⁴ vivo, China

Abstract. Diffusion Transformers (DiTs) have significantly advanced audio-driven portrait animation, but their high computational cost leads to substantial inference latency. Although training-free diffusion caching accelerates inference significant, existing methods are primarily developed for text-conditioned generation and overlook the spatial and modality imbalances inherent in audio-driven portrait animation. In this paper, we propose **SyncCache**, a training-free caching acceleration method tailored for DiT-based portrait animation that explicitly exploits asymmetric dynamics. Specifically, high-frequency dynamics driven by audio conditions and concentrated in human regions are more challenging and critical to cache and reuse than the low-frequency visual background in portrait animation. First, we introduce Spatially-Asymmetric Probing to prioritize error sensitivity in dynamic human region. Second, through Modality-Decoupled Caching, we bypass heavy DiT block by reusing stable inter-block residuals, while continuously recomputing lightweight audio blocks to preserve precise lip synchronization. Furthermore, we introduce a cache ratio to control cache capacity and formulate memory-adaptive cache selection as an offline dynamic programming problem without online overhead. Extensive experiments demonstrate that SyncCache achieves superior speed-quality trade-offs, delivering up to 4.12× acceleration on HunyuanVideo-Avatar and 3.75× on Wan-S2V with near-lossless visual fidelity and precise audio alignment.

Keywords: Diffusion Caching · Audio-driven Portrait Animation · Inference Acceleration

1 Introduction

Diffusion Transformers (DiTs) [29] have advanced video generation, demonstrating exceptional scalability and visual fidelity. Recently, this architecture have

[✉]Corresponding author.

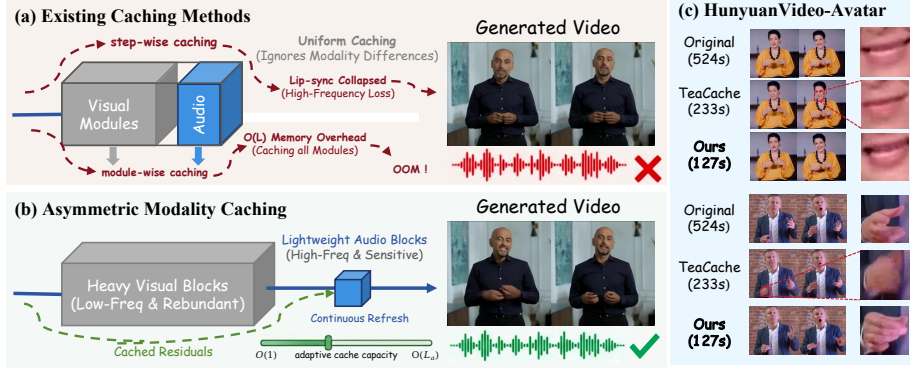


Fig. 1: Comparison of Caching Paradigms. (a) Existing methods blindly assume uniform modality dynamics, leading to either catastrophic high-frequency loss or severe memory overhead. (b) SyncCache explicitly exploits asymmetric modality dynamics, bypassing heavy visual computations via stable residuals while continuously refreshing lightweight audio blocks. (c) Consequently, SyncCache achieves massive acceleration (e.g., $4.12\times$) while preserving delicate high-frequency details that collapse in baselines.

been extended to audio-driven portrait animation [7, 11, 34], enabling the synthesis of realistic talking-face videos from an identity reference and an audio clip. Despite remarkable advances, inference latency remains a critical bottleneck. Generating a short clip requires several minutes even on multi-GPU systems (e.g., generating a 15-second video with HunyuanVideo-Avatar [3] takes approximately 10 minutes on $8\times A800$), severely hindering practical creative iteration.

To accelerate diffusion inference, training-free feature caching has emerged as a prominent solution [5, 9, 21, 23, 26]. However, traditional caching strategies are primarily tailored for text-to-video generation, where the text prompt serves as a low-frequency, global condition. In audio-driven portrait animation, the modality dynamics are highly asymmetric. The reference image provides a low-frequency, static visual prior for the background and identity, whereas the driving audio acts as a high-frequency, local condition that dictates rapid, frame-by-frame lip and muscle movements. Blind to this asymmetry, existing methods exhibit limitations in two distinct ways, as shown in Fig. 1 (a). Timestep-level caching paradigms [21, 26, 46] cache and skip entire denoising steps under the assumption of synchronous modality dynamics, which preserves low-frequency visual content but interrupts the continuous injection of high-frequency audio signals at high acceleration rates. As demonstrated in Fig. 1 (c), this global skipping results in a severe loss of high-frequency details. Conversely, module-level caching [23, 24, 45] attempts to preserve these details by independently caching the feature maps of all modules. This indiscriminate strategy inflates the memory footprint from $O(1)$ to $O(L)$ (where L denotes the number of layers), rapidly exhausting available VRAM. As illustrated in Fig. 2 (a), increasing the

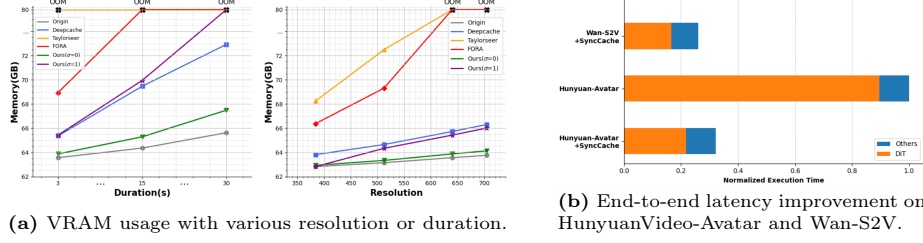


Fig. 2: System-level Efficiency of SyncCache. (a) Unlike existing module-wise methods that rapidly exceed memory constraints as the scale of generation increases, SyncCache maintains a stable and predictable memory footprint. (b) SyncCache significantly reduces the end-to-end inference latency across diverse architectures.

target resolution and video duration causes the memory usage of TaylorSeer [23] and FoRA [30] to increase rapidly and exceed memory constraints.

To resolve this dilemma and accommodate asymmetric dynamics, we propose **SyncCache**, a training-free caching paradigm tailored for audio-driven portrait animation. Instead of uniform step caching or indiscriminate module caching, SyncCache explicitly exploits the dynamic asymmetry by decoupling the computation along both spatial distributions and modality streams.

Specifically, we first leverage the inherent task prior of portrait animation, in which visual dynamics are fundamentally human-centric. This characteristic naturally introduces a spatial asymmetry, where critical high-frequency variations are heavily concentrated on the human subject, whereas the background remains a low-frequency, static anchor that maintains consistency with the reference image. To effectively leverage this prior, we introduce *Spatially-Asymmetric Probing*. Rather than treating the cache errors uniformly across the spatial layout, we weight the probing error with a human mask. This mechanism strategically amplifies the probing sensitivity in audio-aligned, high-frequency regions. Consequently, the model is forced to trigger timely recomputations to preserve and enhance delicate human motions.

To address modality asymmetry, we introduce *Modality-Decoupled Caching*. We observe a severe computational imbalance inherent to portrait animation models: the self-attention and multilayer perceptron (MLP) layers dominate the overall computation, whereas audio blocks are lightweight but essential for injecting high-frequency signals required for precise synchronization and motion. Furthermore, our analysis reveals that the inter-block residuals bridging these audio modules exhibit temporal stability, effectively capturing the slow evolution of features across DiT blocks. Guided by this computational and dynamic asymmetry, we structurally decouple the heavy visual backbone from the audio blocks. We cache the temporally stable inter-block residuals to bypass heavy visual computations, while continuously recomputing lightweight audio blocks.

This ensures the high-frequency control signals remain perfectly synchronized with negligible computational overhead.

Furthermore, to enable highly flexible deployment, SyncCache incorporates a *Memory-Adaptive Optimal Selection* strategy that scales the caching memory overhead from $O(L_a)$ (where L_a denotes the number of audio blocks) towards $O(1)$. We introduce a continuous cache ratio σ to control the cache capacity. For long video generation or environments operating under strict VRAM budgets, setting a smaller σ limits the cache to a specific subset of residual boundaries. We also discover that the temporal stability of inter-block residuals fluctuates significantly across different layers; thus, random or uniform boundary selection is inherently suboptimal and unstable. Therefore, we formulate cache selection as a dynamic programming problem to analytically identify the optimal subset of blocks that strictly minimizes inter-block temporal instability. Crucially, we observe that for any given model, this optimal subset exhibits remarkable robustness and consistency across diverse input samples. Consequently, a single, offline calibration forward pass is sufficient to map out the optimal cache plan for any σ . This zero-overhead calibration guarantees maximum generation fidelity tailored to any specific memory constraint.

Our key contributions are summarized as follows

- *Pioneering Multimodal Caching*: We identify the fundamental limitations of existing caching methods in portrait animation and propose SyncCache, the first cache acceleration method tailored for DiT-based portrait animation that explicitly exploits asymmetric dynamics.
- *Dual-Asymmetry Decoupling*: We design *Spatially-Asymmetric Probing* to amplify sensitivity in dynamic human regions, and *Modality-Decoupled Caching* to continuously refresh high-frequency audio conditions, preserving visual fidelity and flawless lip-sync under extreme speedups.
- *Memory-Adaptive Caching*: We introduce a continuous cache ratio σ to control the cache capacity to enable memory-adaptive caching. By formulating cache block selection as a dynamic programming problem through a single offline calibration, it guarantees maximum temporal stability for any given VRAM constraint without introducing online computational overhead.
- *State-of-the-Art Performance*: Extensive empirical evaluations demonstrate that SyncCache achieves a superior speed-quality trade-off. It delivers up to $4.12\times$ acceleration on HunyuanVideo-Avatar and $3.75\times$ on Wan-S2V while maintaining high-quality generation and audio consistency.

2 Related Works

2.1 Audio-driven Portrait Animation

Portrait animation aims to synthesize a talking-face video from a reference portrait and an input audio clip [11, 31]. Early methods relied on hand-crafted intermediate representations and multi-stage pipelines [37]. With the advance of diffusion models, end-to-end generation has become mainstream. Hallo [40] and

several other methods [4, 16, 28] introduce audio conditioning into U-Net-based pretrained text-to-video models by incorporating audio cross-attention, and they leverage a reference net to preserve identity consistency. However, because U-Net-based diffusion models exhibit limited capacity, the introduction of scalable DiT [29] architectures have significantly accelerated progress in the field of portrait animation. FantasyTalking [34] and Hallo3 [7] explore to condition strong pretrained DiT-based video generation models on audio via cross-modality attention, whereas MultiTalk [18] investigates audio-driven video generation in multi-speaker settings. HunyuanVideo-Avatar [3] is a powerful open-source model that produces dynamic, emotion-controllable, multi-character dialogue videos. WanS2V [12] targets film and television production and delivers realistic visuals, including natural facial expressions, coordinated body motion, and professional cinematography. Despite these advancements associated with the DiT architecture, inference speed remains a significant bottleneck. In this paper, we investigate training-free acceleration during inference and achieve substantial speedups while maintaining near-lossless fidelity.

2.2 Diffusion Model Acceleration

Diffusion models achieve high visual generation quality [17, 33], yet inference latency increases substantially with model capacity and architectural complexity, which hinders deployment in real applications. To address this issue, a broad set of acceleration techniques has been developed, including efficient attention [38, 39, 42], model distillation [27, 35], quantization [10, 44], improved samplers [25], and diffusion cache [2, 5, 22, 41]. Training based approaches often incur additional computational overhead and require auxiliary data, limiting their practical applicability. Conversely, diffusion caching provides a training-free inference acceleration technique by exploiting feature redundancy across adjacent time steps during the sampling process of diffusion models.

Current caching strategies generally fall into two main categories. The first mainstream direction designs decision rules to cache and reuse features at the timestep level. Following this line, TeaCache [21] proposes a polynomial estimator to predict temporal redundancy from input differences. EasyCache [46] replaces this polynomial function with an empirical transformation rate law. MagCache [26] provides a magnitude-aware strategy that adaptively skips timesteps using an error modeling mechanism. DiCache [1] enables the diffusion model to autonomously determine caching timings and adaptively utilize multi-step caches based on an online probe. The second group introduced by TaylorSeer [23] suggests combining multi-step cached features in a Taylor-expansion-like manner to predict reusable features and achieve higher speedups. ClusCa [45] performs spatial clustering on tokens at each timestep, computes only one token per cluster, and propagates this information to all other tokens. SpeCa [24] introduces speculative sampling to diffusion models, predicting intermediate features for subsequent timesteps based on fully computed reference timesteps. Observing that the standard polynomial basis of a Taylor series is suboptimal for model-

ing the complex and non-monotonic trajectories of feature evolution in diffusion models, HiCache [9] proposes a Hermite polynomial-based feature cache.

3 Method

3.1 Preliminary

Flow Matching. Flow Matching [20] formulates generative modeling as deterministic transport governed by an ODE that denoises samples along a straight-line path between a noise prior p_{noise} and the data distribution p_{data} . Given $x_0 \sim p_{\text{data}}$ and $x_T \sim p_{\text{noise}}$, the linear interpolation at timestep $t \in [0, T]$ is

$$x_t = (1 - \frac{t}{T})x_0 + \frac{t}{T}x_T.$$

During training, a timestep-dependent *velocity field* $v_\theta(x_t, t, c)$ is learned to approximate the denoising direction $(x_0 - x_T)$, where c denotes optional conditioning (e.g., text, images, or speech). This yields the neural-ODE dynamics

$$d\hat{x}_t = v_\theta(x_t, t, c) dt.$$

Sampling therefore reduces to integrating the learned velocity field along the straight-line trajectory from p_{noise} to p_{data} .

DiT-based Portrait Animation. Diffusion Transformers [29] have become the dominant backbone for video generation. In portrait animation, the model typically builds upon a pretrained DiT-based image-to-video backbone with interleaving additional audio blocks. These lightweight audio blocks act as high-frequency local controllers, injecting precise lip-sync and human motion signals into the massive backbone.

3.2 SyncCache

To elegantly accommodate the complex dynamics of audio-driven portrait animation, we propose SyncCache, a training-free acceleration paradigm with dual-symmetry decoupling. Instead of uniform step caching or indiscriminate module caching, SyncCache explicitly exploits the dynamic asymmetry inherent to the task by decoupling the computation along both spatial distributions and modality streams. The overall framework is illustrated in Fig. 3.

Spatially-Asymmetric Probing. To accelerate inference via feature caching, it is crucial to estimate the temporal redundancy at each denoising step and determine when to cache. Motivated by prior studies [13, 21], we observe that the feature variation of the first block strongly correlates with the overall denoising dynamics across adjacent timesteps, as shown in Fig. 4. Therefore, we employ the output of the first block, denoted as p^t , as a lightweight probe to dynamically

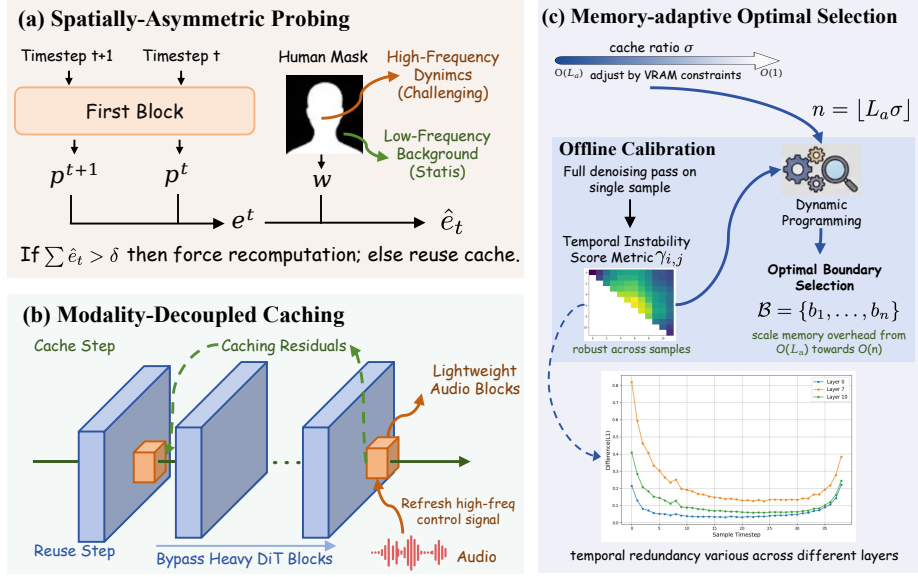


Fig. 3: Overview of the SyncCache framework. SyncCache accelerates audio-driven portrait animation by explicitly exploiting inherent asymmetric dynamics. (a) We utilize a human mask to prioritize computations in highly dynamic human regions. (b) We physically isolate the conditioning stream, bypassing heavy visual DiT blocks via cached residuals while continuously refreshing lightweight audio blocks to maintain flawless lip-sync. (c) We employ an offline calibration and Dynamic Programming (DP) to analytically determine the optimal caching boundaries, gracefully scaling the VRAM footprint according to the cache ratio σ .

determine when to execute a full forward pass and when to safely reuse cache. At each step t , we compute the first block output p^t and measure the relative L1 error with respect to the preceding step as a proxy.

$$e_t = \text{L1}_{\text{rel}}(p^t, p^{t+1}) = \frac{\|p^t - p^{t+1}\|_1}{\|p^{t+1}\|_1}.$$

However, directly relying on a global probing metric overlooks a fundamental task prior: portrait animation exhibits severe spatial asymmetry. Critical high-frequency variations are intensely localized on the human subject, whereas the background remains a low-frequency, static anchor that maintains consistency with the reference image. To strictly enforce the sensitivity of the model to these critical high-frequency details, we modulate the probing error using a spatial human mask M . Crucially, the acquisition of the mask entails minimal computational overhead. It is inherently available as a standard auxiliary input in many DiT-based models [3, 7], and for mask-free architectures [33], M is seamlessly extracted via a lightweight human detector with sub-second latency. The spatially-asymmetric error $\hat{e}_t$ at timestep t is thus formulated as:

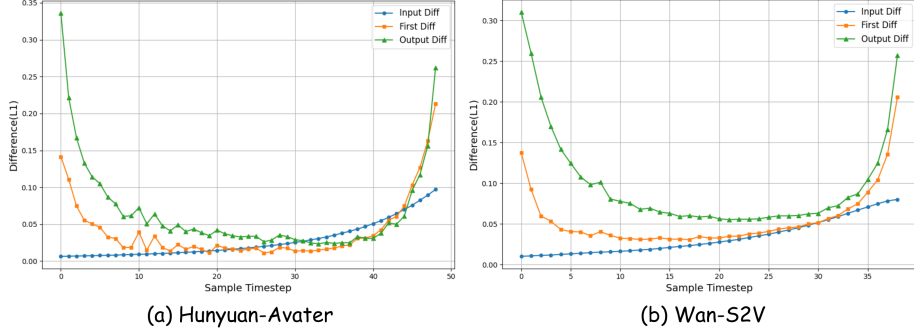


Fig. 4: Visualization of input differences and output differences in consecutive timesteps of HunyuanVideo-Avatar and Wan-S2V. The difference of the first transformer block output exhibits a strong correlation with the difference of the final output between adjacent timesteps in portrait animation models

$$\hat{e}_t = \frac{\| (p^t - p^{t+1}) \odot (1 + \omega M) \|_1}{\| p^{t+1} \odot (1 + \omega M) \|_1},$$

where ω is a human emphasis weight that amplifies probing sensitivity in human regions, and $\odot$ denotes element-wise multiplication.

By strategically tracking this modulated error, we accumulate $\hat{e}_t$ during the cached steps. Once the accumulated error exceeds a user-specified tolerance threshold δ ($\sum \hat{e}_t > \delta$), the model is forced a full forward pass to refresh the cache. This ensures the model instantly triggers recomputations to preserve delicate human motions and triggers timely recomputation to correct error.

Modality-Decoupled Caching. To tackle the inherent modality asymmetry in audio-driven portrait animation, SyncCache introduces a structural separation of the computationally intensive backbone from audio blocks. Within the DiT-based portrait animation pipeline, audio blocks are lightweight and contribute less than 1% to the end-to-end latency, whereas self attention and MLP layers dominate the cost. Despite this low computational overhead, audio blocks are essential for portrait animation and continuously injecting the high-frequency signals required for precise lip synchronization and human motion. Furthermore, as illustrated in Fig. 3 (c), our empirical analysis reveals that the inter-block residuals between these audio modules exhibit high temporal stability across consecutive denoising steps, making them ideal candidates for feature caching.

Guided by this asymmetry, during a timestep with full computation t_a , we cache these stable inter-block residuals $r_{i,j}^{t_a}$ between the i -th and j -th audio block. During subsequent reuse steps, we bypass the computationally expensive DiT blocks by directly reusing the cached residuals $r_{i,j}^{t_a}$. Concurrently, we still compute the lightweight audio blocks. This decoupling ensures that the high-frequency control signals are updated and synchronized at every timestep.

Algorithm 1 Memory-adaptive Caching

-
- 1: **Inputs** number of audio blocks N and cache ratio σ
 - 2: **Output** optimal cache boundaries $\mathcal{B} = \{b_1, \dots, b_n\}$ with $0 = b_0 < b_1 < \dots < b_n = N$
 - 3: **Calibration** run one full denoising pass on a sample and compute $\gamma_{i,j}$ for all $0 \leq i < j \leq N$

$$\gamma_{i,j} = \sum_{t=0}^{T-1} \frac{\|r_{i,j}^t - r_{i,j}^{t+1}\|_1}{\|r_{i,j}^{t+1}\|_1}$$

- 4: $n \leftarrow \lfloor N\sigma \rfloor$
 - 5: $\mathcal{B} \leftarrow \text{CACHESEARCH}(\{\gamma_{i,j}\}, n)$
Execute dynamic programming to select n optimal cache boundaries
 - 6: **Return** $\mathcal{B}$
-

3.3 Memory-adaptive Optimal Selection

Although *Modality-Decoupled Caching* successfully accelerates inference, caching the inter-block residuals at every audio block boundary still scales the memory footprint to $O(L_a)$, where L_a is the total number of audio blocks. For long video generation or deployment on VRAM-constrained devices, this accumulation can rapidly exhaust available memory, as shown in Fig. 2 (a). To enable flexible deployment across diverse hardware constraints, it is essential to dynamically adapt the cache capacity to the specific VRAM budget. Therefore, we introduce a continuous cache ratio $\sigma \in (0, 1]$ that controls the cache capacity to a specific subset of $n = \lfloor L_a \sigma \rfloor$ residual boundaries. However, as shown in Fig. 3 (c), our empirical analysis reveals that the temporal stability of these inter-block residuals varies across different network depths. Consequently, uniform or random selection which boundaries to cache is inherently suboptimal and unstable.

To minimize the global inter-block temporal instability, we formulate this boundary selection as a dynamic programming (DP) problem. We quantify the temporal instability score $\gamma_{i,j}$ for the residual segment between the i -th and j -th audio blocks over the full denoising trajectory T as:

$$\gamma_{i,j} = \sum_{t=0}^{T-1} \frac{\|r_{i,j}^t - r_{i,j}^{t+1}\|_1}{\|r_{i,j}^{t+1}\|_1} \quad (1)$$

Crucially, we observe that for any given model, the optimal path $\mathcal{B}$ exhibits strong robustness across diverse input audio and reference samples. Consequently, a single offline calibration pass is sufficient to determine the optimal cache plan for any specified σ . This guarantees maximum generation fidelity tailored to the VRAM budget, while eliminating online computational overhead.

4 Experiments

4.1 Experimental Settings

We conduct our main experiments on two portrait animation models, Wan-S2V [12] and HunyuanVideo-Avatar [3], to evaluate the effectiveness of our method. HunyuanVideo-Avatar is a portrait animation model built on HunyuanVideo, which injects audio conditioning by inserting audio blocks after the dual-stream DiT blocks. We follow the standard 50-step inference protocol as the baseline and keep all default sampling parameters to ensure strict experimental consistency. Wan-S2V is an audio-driven video generation model built on Wan-14B. It injects audio conditioning by inserting eleven audio blocks after backbone blocks and it uses a 40-step UniPC sampler.

Evaluation and Metrics Following prior work [12, 18, 19] in portrait animation, we perform evaluations on the EMTD dataset [28], which primarily consists of solo-talking and semi-body human videos. To assess the quality of the talking-face videos generated under acceleration, we evaluate three aspects: inference efficiency, visual quality, and audio-visual synchronization. Regarding inference efficiency, we report the speedup ratio and the inference latency. For visual quality, in line with previous studies on caching acceleration [2, 21, 23], we utilize LPIPS [43], PSNR [15], and SSIM [36] to evaluate the fidelity of generated videos relative to origin results. Consistent with the portrait animation methods [4, 31, 40], we additionally employ Frechet Inception Distance (FID) [14] and Frechet Video Distance (FVD) [32] to measure the distance between generated and real videos. To evaluate the alignment between lip movements and audio signals, we use Sync-C and Sync-D [6], metrics that are widely adopted within the portrait animation community.

Implementation Details All experiments are conducted on NVIDIA A800 80GB GPUs with PyTorch and FlashAttention [8] enabled by default. For latency benchmarking, we use $8 \times$ A800 GPUs. We apply the default FSDP configuration to HunyuanVideo-Avatar. For Wan-S2V, we use the default FSDP to shard the DiT and the text encoder together with Ulysses sequence parallelism. We set the human emphasis weight ω to 2 (empirically robust for $\omega \in [2, 4]$). For mask-free models (e.g., Wan-S2V), we employ the same lightweight detection model (containing 46 million parameters) as utilized in Hunyuan-Avatar, which incurs a negligible latency of 0.07 seconds.

4.2 Quantitative Comparison

Performance on HunyuanVideo-Avatar. As reported in Tab. 1, SyncCache delivers the best overall performance, surpassing existing baselines across almost all key metrics. Under a conservative threshold, SyncCache-slow reduces end-to-end latency from nearly 10 minutes to 157 seconds while achieving the best visual fidelity across LPIPS (0.1016), SSIM (0.8618), FID (25.65), and FVD

Table 1: Quantitative comparison with other methods on HunyuanVideo-Avatar. The best value is in **bold** and the second best is underlined. **OOM** indicates a CUDA out-of-memory error on eight A800 GPUs with 80GB memory each. [†]TaylorSeer-series denotes a set of recent methods derived from TaylorSeer.

Method	Visual Quality					Audio Consistency		Acceleration	
	LPIPS↓	PSNR↑	SSIM↑	FID↓	FVD↓	Sync-C↑	Sync-D↓	Speedup↑	Latency(s)↓
Original: 50 steps	-	-	-	25.27	240.16	6.963	8.640	-	524
Δ -DiT [2]	0.1321	23.26	0.8281	26.52	251.68	6.652	8.878	1.38×	381
TeaCache [21]	0.1730	<u>25.55</u>	0.8428	26.83	236.08	6.842	8.712	2.25×	233
MagCache [26]	0.1696	25.76	0.8455	26.43	<u>235.87</u>	6.830	<u>8.664</u>	2.30×	228
DiCache [1]	0.1548	25.29	0.8490	<u>26.12</u>	238.15	6.834	8.726	2.41×	217
TaylorSeer-series [†] [23, 24, 45]	OOM	OOM	OOM	OOM	OOM	OOM	OOM	-	-
CGCache [13]	0.1848	24.05	0.8249	27.75	238.51	6.814	8.755	3.18×	164
SyncCache-slow	0.1016	24.93	0.8618	25.65	234.86	6.944	8.653	3.34×	<u>157</u>
SyncCache-fast	<u>0.1172</u>	24.41	<u>0.8493</u>	26.89	241.27	<u>6.902</u>	8.673	4.12×	127

Table 2: Quantitative comparison with other methods on Wan-S2V. The best value is in **bold** and the second best is underlined. **OOM** indicates a CUDA out-of-memory error on eight A800 GPUs with 80GB memory each. [†]TaylorSeer-series denotes a set of recent methods derived from TaylorSeer.

Method	Visual Quality					Audio Consistency		Acceleration	
	LPIPS↓	PSNR↑	SSIM↑	FID↓	FVD↓	Sync-C↑	Sync-D↓	Speedup↑	Latency (s)↓
Original (40 steps)	-	-	-	36.60	293.05	6.712	8.632	-	113
TeaCache [21]	0.1863	19.04	0.7468	39.45	297.45	6.678	8.642	2.93×	38.62
MagCache [26]	0.1839	<u>19.46</u>	<u>0.7524</u>	32.93	287.36	6.709	8.641	2.96×	38.15
Δ -DiT [2]	0.1869	19.24	0.7445	38.91	277.33	6.693	8.616	1.73×	65.26
CGCache [13]	0.1884	19.20	0.7443	37.17	298.76	6.710	8.651	3.00×	37.65
Taylorseer-series [†] [23, 24, 45]	OOM	OOM	OOM	OOM	OOM	OOM	OOM	-	-
DiCache [1]	<u>0.1835</u>	19.43	0.7498	34.17	281.56	<u>6.716</u>	<u>8.614</u>	2.99×	37.82
SyncCache	0.1775	19.80	0.7665	<u>33.83</u>	<u>280.78</u>	6.791	8.541	3.75×	30.15

(234.86). Most crucially, regarding audio consistency, SyncCache-slow achieves a near-lossless Sync-C of 6.944 (vs. 6.963 Original), markedly higher than existing methods like TeaCache and MagCache. This validates that our strategy of exploiting asymmetric dynamic is essential for preserving delicate high-frequency human dynamics. Pushing to a 4.12× speedup, SyncCache-fast still dominates all baselines in both perceptual quality and lip-sync alignment. In contrast, existing methods reveal intrinsic limitations due to their blind uniform assumption. Timestep-level caching (TeaCache, MagCache) saturates speedups around 2.30× and discards high-frequency audio conditions, leading to a noticeable degradation in Sync-C. TaylorSeer-series exceed memory constraint and trigger CUDA out-of-memory (OOM) failures due to indiscriminate $O(L)$ feature caching.

Performance on Wan-S2V. Wan-S2V presents a more severe challenge due to its extremely compressed 40-step sampler and complex dynamic patterns. However, as demonstrated in Tab. 2, SyncCache effectively adapts to this challenging context, demonstrating substantial improvements over existing baselines. When traditional caching methods are applied to Wan-S2V, the lack of explicit

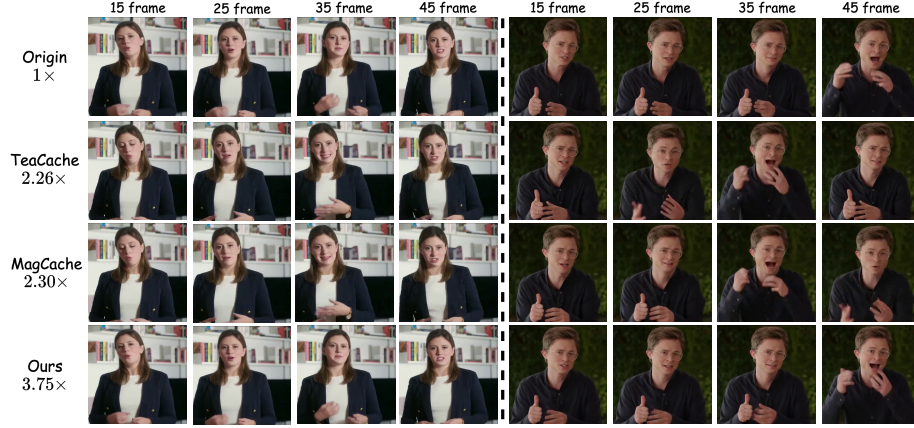


Fig. 5: Comparison of visual quality and efficiency (denoted by speedup ratio) with other methods on HunyuanVideo-Avatar.

modality modeling leads to compromised performance. Both TeaCache and MagCache exhibit observable degradations in audio consistency and visual fidelity. The step-wise skipping mechanism of these methods interrupts the continuous high-frequency audio feed required to stabilize visual motion, which results in noticeable lip-sync misalignment. In contrast, SyncCache successfully addresses this issue by targeting the dual asymmetry in the dynamics of portrait animation. As a result, SyncCache dominates across all quality metrics on Wan-S2V, attaining the best LPIPS (0.1775), PSNR (19.80), and Sync-C (6.791) at $3.75\times$ speedups. This consistent success generalizes the necessity of exploiting asymmetric dynamics to preserve delicate high-frequency motion.

4.3 Qualitative Comparison

We provide visualizations to compare the quality of generated videos against that of several baseline methods. As demonstrated in Fig. 5, our method better preserves high-frequency details and temporal alignment at higher speedup ratios. Traditional caching methods, such as MagCache and TeaCache, operate under synchronous modality dynamics, which leads to the loss of high-frequency details. This degradation manifests as blurry hand gestures and poor lip-synchronization. In contrast, SyncCache achieves notably better lip-audio alignment compared to other approaches at an acceleration factor of $4.12\times$.

Effectiveness in Multi-Person Scenarios. Research on portrait animation is actively expanding to multi-speaker scenarios, often employing a dynamic masking strategy to control the range of audio cross attention constraints [3, 18]. We integrate this masking strategy into HunyuanVideo-Avatar and evaluate the compatibility of the proposed method with this approach. As shown in Fig. 6,

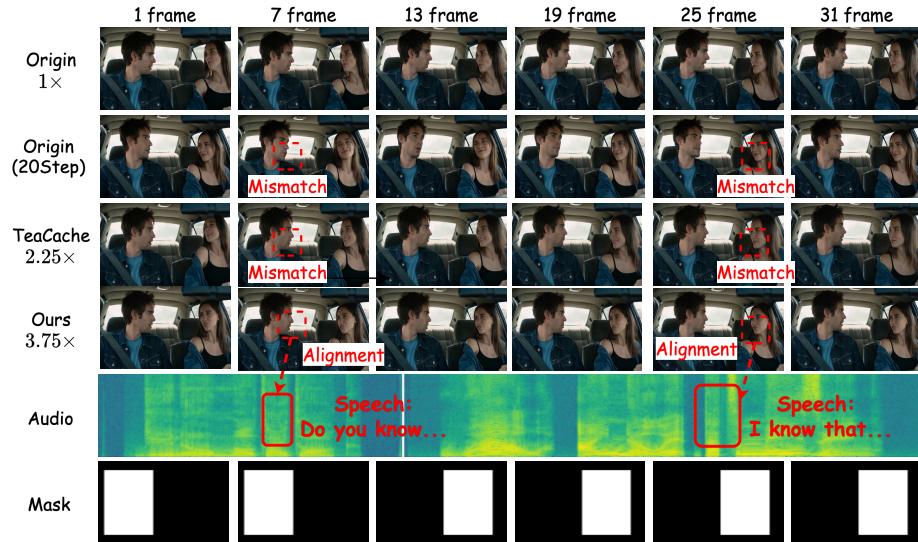


Fig. 6: Qualitative comparison in multi-person scenarios. Compared with the original method and TeaCache [21], our method exhibits stronger consistency for the utterances “Do you know...” and “I know that...” in multi-speaker scenarios.

Table 3: Ablation study of SyncCache components on HunyuanVideo-Avatar.

Method	LPIPS↓	PSNR↑	SSIM↑	Sync-C↑	Sync-D↓	Latency(s)↓
w/o SAP	0.1259	24.76	0.8534	6.867	8.684	161
w/o MDC	0.1571	24.54	0.8437	6.822	8.735	156
Ours	0.1016	24.93	0.8618	6.944	8.653	157

SyncCache produces significantly finer and more accurate lip synchronization details in multi-speaker dialogue scenarios. We attribute this improvement to Modality-Decoupled Caching (MDC). Baseline methods such as TeaCache experience synchronization failures due to the weak control of audio conditions, resulting in misaligned lip movements. In contrast, MDC ensures that the audio condition is reinforced at every step. By continuously refreshing the audio controls and incorporating mask information, MDC enables SyncCache to operate seamlessly with complex spatial masks without sacrificing temporal alignment.

4.4 Ablation Study

Ablation of Core SyncCache Components. To validate the effectiveness of the proposed design, we examine the contributions of Spatially-Asymmetric Probing (SAP) and Modality-Decoupled Caching (MDC).

Table 4: Ablation study of Memory-adaptive Optimal Selection on HunyuanVideo-Avatar with $\sigma = 0.4$.

Method	LPIPS↓	PSNR↑	SSIM↑	Sync-C↑	Sync-D↓
1 Sample	0.1016	24.93	0.8618	6.944	8.653
All Samples	0.0981	25.01	0.8603	6.951	8.645
Silent Sample	0.1021	24.94	0.8608	6.932	8.671
w/o DP	0.1369	24.69	0.8526	6.847	8.709

Effect of Spatially-Asymmetric Probing. Removing SAP (i.e., reverting to a naive, uniform probing metric) leads to a clear degradation in spatial visual fidelity. As demonstrated in Tab. 3, all metrics noticeably deteriorates without SAP. This empirical drop confirms the initial intuition: without explicitly prioritizing highly dynamic human regions, the probing mechanism suffers from global dilution. Consequently, it prematurely skips computations and fails to preserve intricate, high-frequency human details. In cases where the confidence of the human mask is low or the detection fails, our pipeline naturally defaults to using the full image. Crucially, Spatially-Asymmetric Probing utilizes a soft multiplicative mask rather than a hard crop. Therefore, an imperfect mask simply causes the model to degrade gracefully toward uniform probing, rather than imposing incorrect spatial asymmetries.

Effect of Modality-Decoupled Caching. Disabling MDC equates to a rigid, synchronous caching strategy where both visual and audio blocks are skipped simultaneously. This omission causes a significant degradation in temporal alignment, with the Sync-C metric decreasing from 6.944 to 6.822. Crucially, a comparison between the configuration without MDC and the proposed approach reveals that restoring the continuous recomputation of audio blocks recovers near-lossless lip-sync accuracy with a slight latency penalty. This finding demonstrates that isolating and continuously refreshing the lightweight audio stream is a necessity with zero overhead for preserving high-frequency synchronization.

Effectiveness and Robustness of Memory-adaptive Optimal Selection.

We evaluate the Memory-adaptive Optimal Selection strategy regulated by dynamic programming (DP) to verify the effectiveness and robustness of the method across diverse samples. As demonstrated in Tab. 4, analytically identifying the optimal subset of boundaries based on residual stability achieves superior generalization compared to arbitrary selection. Compared to the arbitrary selection setting (w/o DP), the proposed DP-based strategy improves performance, reducing the LPIPS from 0.1369 to 0.1016 and increasing the Sync-C from 6.847 to 6.944. Furthermore, the robustness and consistency of this optimal path remain stable across different calibration scenarios. We observe negligible performance differences when optimizing the cache path based on a single sample instead of the average of all samples. Even utilizing a silent sample with minimal lip motion

to determine the cache boundaries yields results comparable to those obtained from active talking samples. This result indicates that a single offline calibration pass can satisfy specific memory constraints without introducing online computational overhead. These findings confirm that optimal cache boundaries depend on the inter-layer temporal dynamics of the model rather than specific input prompts, thereby validating the robustness of the proposed calibration method.

5 Conclusion

We propose SyncCache, a training-free, memory-adaptive caching strategy tailored for DiT-based audio-driven portrait animation. By explicitly exploiting the asymmetric dynamics of the generation process, we developed Spatially-Asymmetric Probing to prioritize high-frequency human motion regions and Modality-Decoupled Caching to continuously refresh lightweight audio conditions while bypassing heavy, redundant visual computations. Furthermore, we incorporated a memory-adaptive optimal selection strategy that leverages offline dynamic programming to dynamically scale cache capacity according to VRAM constraints without incurring online overhead. Experiments show that SyncCache achieves a superior speed–quality trade-off, delivering a $4.12\times$ acceleration on HunyuanVideo-avatar and $3.75\times$ on Wan-S2V with negligible degradation in visual quality or audio alignment.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China (No. 62332002, 62425101), The Guangdong Grants (Grant No.2023ZT10X075), and Shenzhen KQTD (No.20240729102051063).

References

1. Bu, J., Ling, P., Zhou, Y., Wang, Y., Zang, Y., Lin, D., Wang, J.: Dicache: Let diffusion model determine its own cache. In: The Fourteenth International Conference on Learning Representations (2026)
2. Chen, P., Shen, M., Ye, P., Cao, J., Tu, C., Bouganis, C.S., Zhao, Y., Chen, T.: δ -dit: Accelerating diffusion transformers without training via denoising property alignment. *International Journal of Computer Vision* **134**(6), 276 (2026)
3. Chen, Y., Liang, S., Zhou, Z., Huang, Z., Ma, Y., Tang, J., Lin, Q., Zhou, Y., Lu, Q.: Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. *arXiv preprint arXiv:2505.20156* (2025)
4. Chen, Z., Cao, J., Chen, Z., Li, Y., Ma, C.: Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 2403–2410 (2025)
5. Chu, H., Wu, W., Feng, G., Zhang, Y.: Omnicache: A trajectory-oriented global perspective on training-free cache reuse for diffusion transformer models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 16302–16312 (2025)

6. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Asian conference on computer vision. pp. 251–263. Springer (2016)
7. Cui, J., Li, H., Zhan, Y., Shang, H., Cheng, K., Ma, Y., Mu, S., Zhou, H., Wang, J., Zhu, S.: Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21086–21095 (2025)
8. 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)
9. Feng, L., Zheng, S., Liu, J., Lin, Y., Zhou, Q., Cai, P., Wang, X., Chen, J., Zou, C., Ma, Y., et al.: Hicache: Training-free acceleration of diffusion models via hermite polynomial-based feature caching. *arXiv preprint arXiv:2508.16984* (2025)
10. Feng, W., Yang, C., Qin, H., Li, X., Wang, Y., An, Z., Huang, L., Diao, B., Zhao, Z., Xu, Y., et al.: Q-vdit: Towards accurate quantization and distillation of video-generation diffusion transformers. *arXiv preprint arXiv:2505.22167* (2025)
11. Gan, Q., Yang, R., Zhu, J., Xue, S., Hoi, S.: Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. *arXiv preprint arXiv:2506.18866* (2025)
12. Gao, X., Hu, L., Hu, S., Huang, M., Ji, C., Meng, D., Qi, J., Qiao, P., Shen, Z., Song, Y., et al.: Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621* (2025)
13. Guan, X., Jiang, L., Chen, H., Zhang, X., Yan, J., Wang, G., Liu, Y., Zhang, Z., Wu, Y.: Forecasting when to forecast: Accelerating diffusion models with confidence-gated taylor. *Knowledge-Based Systems* p. 114635 (2025)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
15. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: 2010 20th international conference on pattern recognition. pp. 2366–2369. IEEE (2010)
16. Jiang, J., Liang, C., Yang, J., Lin, G., Zhong, T., Zheng, Y.: Loopy: Taming audio-driven portrait avatar with long-term motion dependency. In: The Thirteenth International Conference on Learning Representations (2025)
17. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024)
18. Kong, Z., Gao, F., Zhang, Y., Kang, Z., Wei, X., Cai, X., Chen, G., Luo, W.: Let them talk: Audio-driven multi-person conversational video generation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
19. Li, X., Xie, P., Ren, Y., Gan, Q., Zhang, C., Kong, F., Yin, X., Peng, B., Yuan, Z.: Infinityhuman: Towards long-term audio-driven human. *arXiv preprint arXiv:2508.20210* (2025)
20. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022)
21. Liu, F., Zhang, S., Wang, X., Wei, Y., Qiu, H., Zhao, Y., Zhang, Y., Ye, Q., Wan, F.: Timestep embedding tells: It’s time to cache for video diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7353–7363 (2025)
22. Liu, J., Cai, P., Zhou, Q., Lin, Y., Kong, D., Huang, B., Pan, Y., Xu, H., Zou, C., Tang, J., et al.: Freqca: Accelerating diffusion models via frequency-aware caching. *arXiv preprint arXiv:2510.08669* (2025)

23. Liu, J., Zou, C., Lyu, Y., Chen, J., Zhang, L.: From reusing to forecasting: Accelerating diffusion models with taylorseers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15853–15863 (2025)
24. Liu, J., Zou, C., Lyu, Y., Ren, F., Wang, S., Li, K., Zhang, L.: Spec: Accelerating diffusion transformers with speculative feature caching. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 10024–10033 (2025)
25. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems* **35**, 5775–5787 (2022)
26. Ma, Z., Wei, L., Wang, F., Zhang, S., Tian, Q.: Magcache: Fast video generation with magnitude-aware cache. *Advances in Neural Information Processing Systems* **38**, 34348–34380 (2026)
27. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14297–14306 (2023)
28. Meng, R., Zhang, X., Li, Y., Ma, C.: Echomimicv2: Towards striking, simplified, and semi-body human animation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5489–5498 (2025)
29. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
30. Selvaraju, P., Ding, T., Chen, T., Zharkov, I., Liang, L.: Fora: Fast-forward caching in diffusion transformer acceleration. *arXiv preprint arXiv:2407.01425* (2024)
31. Tian, L., Wang, Q., Zhang, B., Bo, L.: Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In: European Conference on Computer Vision. pp. 244–260. Springer (2024)
32. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation (2019)
33. 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)
34. Wang, M., Wang, Q., Jiang, F., Fan, Y., Zhang, Y., Qi, Y., Zhao, K., Xu, M.: Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 9891–9900 (2025)
35. Wang, X., Zhang, S., Zhang, H., Liu, Y., Zhang, Y., Gao, C., Sang, N.: Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109* (2023)
36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)
37. Wei, H., Yang, Z., Wang, Z.: Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694* (2024)
38. Xi, H., Yang, S., Zhao, Y., Xu, C., Li, M., Li, X., Lin, Y., Cai, H., Zhang, J., Li, D., et al.: Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. *arXiv preprint arXiv:2502.01776* (2025)
39. Xia, Y., Ling, S., Fu, F., Wang, Y., Li, H., Xiao, X., Cui, B.: Training-free and adaptive sparse attention for efficient long video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15982–15993 (2025)
40. Xu, M., Li, H., Su, Q., Shang, H., Zhang, L., Liu, C., Wang, J., Yao, Y., Zhu, S.: Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801* (2024)

41. Zhang, H., Gao, T., Shao, J., Wu, Z.: Blockdance: Reuse structurally similar spatio-temporal features to accelerate diffusion transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12891–12900 (2025)
42. Zhang, J., Xiang, C., Huang, H., Xi, H., Zhu, J., Chen, J., et al.: Spargeattention: Accurate and training-free sparse attention accelerating any model inference. In: Forty-second International Conference on Machine Learning (2025)
43. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
44. Zhao, T., Fang, T., Huang, H., Wan, R., Soedarmadji, W., Liu, E., Li, S., Lin, Z., Dai, G., Yan, S., et al.: Vedit-q: Efficient and accurate quantization of diffusion transformers for image and video generation. In: The Thirteenth International Conference on Learning Representations (2025)
45. Zheng, Z., Wang, X., Zou, C., Wang, S., Zhang, L.: Compute only 16 tokens in one timestep: Accelerating diffusion transformers with cluster-driven feature caching. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 10181–10189 (2025)
46. Zhou, X., Liang, D., Chen, K., Feng, T., Chen, X., Lin, H., Ding, Y., Tan, F., Zhao, H., Bai, X.: Less is enough: Training-free video diffusion acceleration via runtime-adaptive caching. arXiv preprint arXiv:2507.02860 (2025)