DYNAMIC ATTENTION DECOMPOSITION FOR EFFICIENT VIDEO DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion Transformers (DiTs) have achieved remarkable success in video generation, but their core Multi-Head Attention (MHA) mechanism has become the primary bottleneck for inference efficiency due to significant computational redundancy. Existing acceleration methods either perform an overly simplistic hard partitioning of attention heads or suffer from deficiencies in efficiency and flexibility in their combination methods. Established eigenvalue analysis has confirmed that the behavior of attention heads exists on a low-dimensional manifold, indicating a high degree of functional redundancy and providing a theoretical basis for our decomposition method. Based on this recognized redundancy, we propose a novel inference acceleration framework: Dynamic Attention Decomposition (DAD). DAD decomposes the output of any attention head into a "soft" weighted combination of a fixed, globally shared set of universal spatio-temporal bases. To achieve this input-adaptive combination, we designed an anchor direction projection mechanism that computes the optimal weights in real-time based on input features, with negligible additional overhead. Furthermore, to address the disruption to Rotary Positional Encoding (RoPE) caused by the decomposition and to compensate for approximation errors, we propose a novel dual residual module. By decoupling the tasks, it separately restores high-frequency positional signals and compensates for the remaining low-frequency content. Experiments demonstrate that our method achieves significant end-to-end acceleration while maintaining high-quality generation, providing a new pathway for the efficient deployment of large-scale video generation models.

1 INTRODUCTION

In recent years, Diffusion Transformers (DiTs) have become a transformative paradigm in video generation, achieving unprecedented success in generating high-fidelity and temporally coherent videos by extending the self-attention mechanism to 3D spacetime. However, the immense computational overhead—especially the 3D attention, which accounts for over 60% of the total latency—severely hinders the practical deployment of these models (Xi et al., 2025). This computational bottleneck, which scales quadratically with context length, becomes increasingly dominant as video resolution and duration increase.

It has long been recognized in the academic community that the Multi-Head Attention (MHA) mechanism exhibits significant redundancy. Seminal eigenvalue analysis, particularly in language models, has confirmed that the behavior of numerous attention heads actually lies on a shared **low-dimensional manifold**, where a small number of principal components expand the vast majority of the variance (Nguyen et al., 2022; Hu et al., 2024; Zhou et al., 2025). This finding indicates that the "effective" number of heads is far smaller than the nominal number, providing a solid theoretical foundation for decomposing complex attention computations onto a shared, more compact set of bases.

Building on this recognized redundancy, existing acceleration solutions either enforce a hard partitioning of attention heads (e.g., classifying heads as strictly "spatial" or "temporal") or gravitate towards two extremes (Xi et al., 2025; Sun et al., 2025; Yang et al., 2025). The first is **top-level late mixing**, which applies a dynamic selection or weighting to the outputs of the heads only after each independent head has already completed its expensive attention computation (Ahmed et al.,

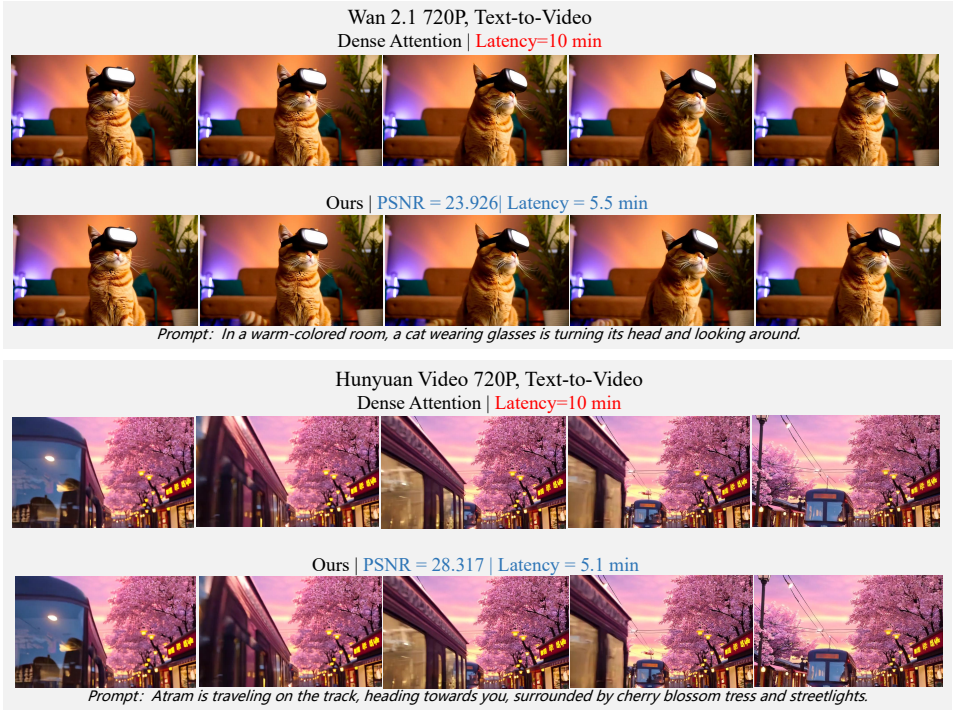


Figure 1: Our proposed DAD method accelerates video generation while maintaining high quality. On a single A100, for Hunyuan Video and Wan 2.1, DAD achieves up to 1.96× and 1.82× end-to-end speedup, with a PSNR up to 28 and 24.

2017; Li et al., 2019; Zhang et al., 2022). The second is **bottom-level projection composition** (Cordonnier et al., 2020; Liu et al., 2022; Kazemnejad et al., 2023), which reassembles the $Q/K/V$ projection matrices in the linear layer to "generate new heads." The former approach composes too late, leaving the token-to-token information flow within the attention stage unchanged. The latter, while able to fundamentally alter the information flow, is typically static and incurs considerable extra computational and implementation complexity to form the new combined projections. Meanwhile, our statistical analysis (Figure 2) reveals that a large number of attention heads exhibit mixed spatio-temporal properties rather than a clean dichotomy, further limiting the effectiveness of hard partitioning schemes.

To this end, we propose Dynamic Attention Decomposition (DAD): an input-adaptive soft combination at the attention output Y level. Specifically, the output Y of any given head is expressed as a weighted sum of the outputs from a globally shared set of spatio-temporal bases, supplemented by a lightweight low-rank residual to correct for approximation errors. The weights are generated by an anchor direction projection mechanism, which treats the current head's input feature as an "anchor direction" and dynamically produces combination coefficients based on its alignment with each basis output. In this way, we first compute the small set of shared bases (once), and then perform an extremely lightweight combination and compensation for each head. Compared to top-level late mixing, our method alters the combination during the attention stage itself; compared to bottom-level projection composition, it maintains computational and implementation simplicity (with almost no additional overhead) and is adaptive to the input.

However, this low-rank decomposition at the output level can disrupt the vector structure relied upon by Rotary Positional Encoding (RoPE). To address this challenge and further enhance approximation accuracy, we designed an innovative dual residual module. This module bifurcates the compensation task: one residual, R_{pos}, focuses on restoring the high-frequency positional signals primarily contributed by RoPE, while the other, R_{free}, under an orthogonality constraint, compensates for the remaining low-frequency content not covered by the shared bases. This decomposition shifts the computational bottleneck from numerous expensive, independent full-attention operations to just a few efficient, globally shareable basis operations, thereby achieving significant inference acceleration, while simultaneously preserving the fidelity of generated results through the two low-rank residuals, as shown in Figure 1.

Our main contributions are summarized as follows:

- We propose a unified framework, Dynamic Attention Decomposition (DAD), that performs dynamic decomposition at the attention output level, and we design a novel anchor direction projection mechanism that achieves efficient, adaptive approximation of different head characteristics with negligible additional overhead.
- We design a new dual residual module that, through a separation of concerns, structurally resolves the two key challenges introduced by low-rank approximation: positional encoding integrity and generative diversity.
- DAD accelerates video generation while maintaining high quality. On a single NVIDIA A100 GPU, for Hunyuan Video and Wan 2.1, DAD achieves up to 1.96× and 1.82× end-to-end speedup, respectively, with PSNR scores reaching up to 28 and 24.

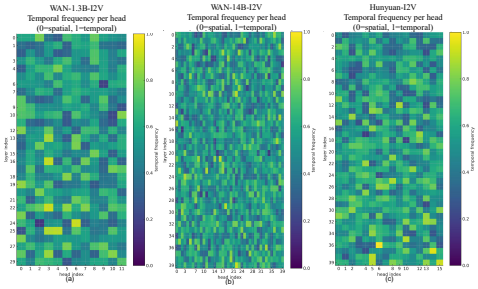


Figure 2: Head-wise temporal frequency heatmaps (0 = spatial, 1 = temporal) for three open-source video generators: (a) WAN-2.1-1.3B, (b) WAN-2.1-14B, (c) Hunyuan. A large portion of heads lie in the 0.3-0.7 "mixed band," indicating pervasive spatio-temporal mixed heads rather than a simple spatial/temporal dichotomy. This motivates our anchor-projection routing that softly combines shared bases instead of hard head classification.

2 Realted Work

Diffusion Model Inference Acceleration. Existing acceleration methods for diffusion models primarily focus on two strategies. The first category reduces the number of denoising steps using techniques like adjusted samplers or distillation (Yin et al., 2025; Li et al., 2024a), but these approaches are often sensitive to hyperparameters or require costly retraining, which can degrade generation quality. The second category exploits temporal redundancy by caching features such as attention maps across steps (Ma et al., 2024; Liu et al., 2025; Zhao et al., 2024a), yet this can also compromise fidelity. Our DAD framework is distinct, as it focuses on improving the core attention mechanism itself. As an orthogonal approach, the efficiency gains from DAD can be combined with other methods, such as caching, to achieve even greater speedups.

Efficient Attention Mechanisms. Other methods aim to improve the efficiency of the attention mechanism itself (Zhang et al., 2025b;a; Xu et al., 2025; Xia et al., 2025; Xi et al., 2025). Linear attention (Shen et al., 2021; Katharopoulos et al., 2020; Han et al., 2024; Wang et al., 2020) achieves linear complexity but often requires architectural changes and can underperform compared to full attention. Sparse attention methods (Li et al., 2025a;b; Lai et al., 2025; Xia et al., 2025; Li et al., 2025b; Yang et al., 2025) limit each token's attention scope, but for video, these methods often rely on heuristics, failing to address the redundancy across heads. In contrast, DAD introduces a new perspective by decomposing attention onto a universal spatiotemporal basis. Instead of finding sparse patterns, DAD projects attention onto a shared set of bases, thereby resolving computational redundancy. This decomposition enables real-time, content-aware weighting and, through its multifunctional residual module, systematically maintains positional encoding integrity and generative diversity—key challenges often overlooked by most existing methods.

3 Methodology

To achieve efficient and high-quality inference acceleration for DiT models, we propose the DAD framework, as shown in Figure3. This framework significantly reduces the computational complexity of inference while preserving generation quality by decomposing the compute-intensive full

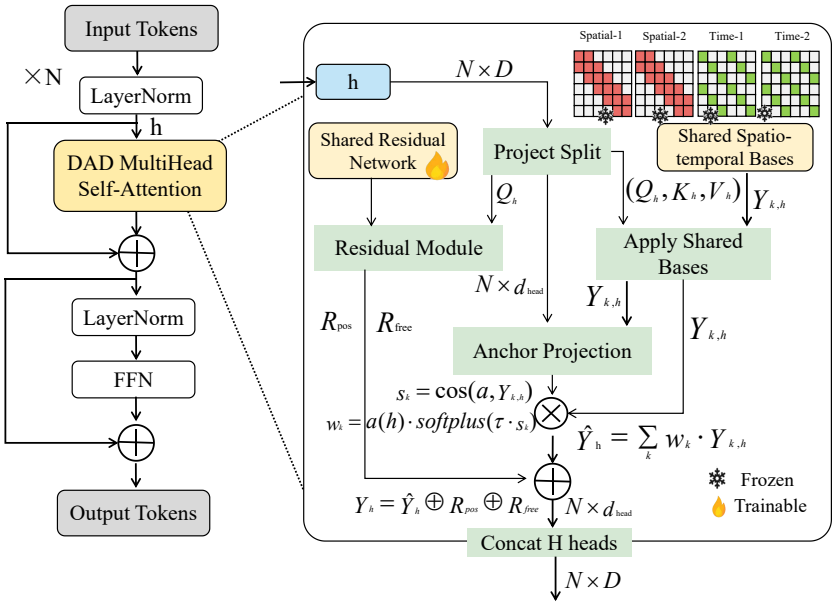


Figure 3: **DAD Framework.** A small set of shared spatio-temporal bases is computed once per layer and reused across heads. For each head, anchor-direction projection in value space scores each base by cosine similarity $s_k = \cos(a, Y_{k,h})$, and produces *decoupled* weights without a simplex constraint: $\omega_k = a(h)\,\mathrm{softplus}(\tau\, s_k)$. The head output is $\hat{Y}_h = \sum_k \omega_k\, Y_{k,h} + R_{\text{pos}} + R_{\text{free}}$, where a shared micro-network predicts low-rank residuals R_{pos} (positional/high-frequency restoration) and R_{free} (content backfilling). Heavy computation is concentrated in the shared bases; per-head cost remains lightweight, and heads are concatenated back to $N \times D$.

attention operation into lightweight operations based on a set of shared bases. This section provides a detailed exposition of its core design and implementation.

3.1 Framework Overview and Core Insight

The remarkable performance of modern Video Diffusion Models (DiTs) heavily relies on their stacked multi-head self-attention modules. However, this design comes at the cost of an $O(N^2)$ computational complexity, where N is the total number of spatio-temporal tokens. Through an in-depth analysis of these models, we discovered a high degree of functional overlap among different attention heads, with many heads sharing similar spatio-temporal information processing patterns. This observation reveals a prevalent computational redundancy in existing architectures.

Our core insight is that the complex function of any attention head can be approximated as a linear projection onto a shared, more fundamental set of **Spatio-temporal Bases**. These bases act as "knowledge genes," encapsulating the model's most critical priors for processing spatio-temporal information. Based on this insight, the DAD framework deconstructs the conventional single-step full attention computation into a three-stage synergistic process:

1. **Basis Decomposition:** First, we approximate the computation of an arbitrary attention head as a weighted combination of projections onto a fixed, shared set composed of multiple temporal and spatial bases.
2. **Dynamic Weighting:** Second, we introduce a novel **Anchor Direction Projection** mechanism to dynamically and content-awaresly compute the combination weights for each basis, conditioned on the real-time input features.
3. **Residual Compensation:** Finally, we design a multifunctional residual module that not only corrects for the approximation errors introduced by the decomposition but is also endowed with the advanced functions of compensating for critical positional information and injecting generative diversity.

This process centralizes the heavy computation into a one-time, shareable basis calculation, while realizing the unique function of each head through extremely lightweight weighting and residual correction, thereby striking a balance between efficiency and performance.

3.2 Unified Spatio-temporal Basis: Selection, Construction, and Decomposition

The cornerstone of the DAD framework is the concept of a shared spatio-temporal basis. To ensure this set of bases is both compact and sufficiently expressive, we employ a data-driven approach for its selection and construction.

3.2.1 Basis Selection and Construction: An Offline Clustering Analysis

Our basis set is not determined empirically but is derived from an in-depth offline analysis of a pretrained DiT model. We first run the model on a diverse video dataset, collecting the attention maps (QK^T matrices) from all attention heads across different layers and timesteps. Subsequently, by performing K-Means clustering on this vast collection of vectorized attention maps, we find that the attention patterns clearly converge into several major categories.

The two most dominant categories are Spatial Locality and Temporal Correlation. Finer-grained clustering reveals sub-patterns within each major category; for instance, spatial patterns can be subdivided into "texture details" and "object contours," while temporal patterns can be split into "smooth motion" and "abrupt changes", as shown in Figure 4.

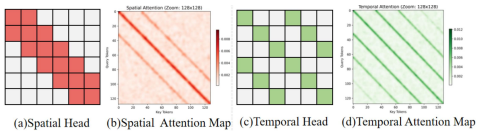


Figure 4: (a,b) Spatial Heads focus on intra-frame tokens (red), while (c,d) Temporal Heads attend to tokens at the same spatial location across frames (green).

Based on this discovery, we determined that a configuration of $M = 4$ **temporal bases** and $N = 4$ **spatial bases** offers an optimal trade-off between expressiveness and efficiency. These four bases are instantiated as independent attention modules, whose parameters are optimized to precisely replicate the centroid patterns of their corresponding clusters and are subsequently **parameter-frozen** within the DAD framework.

3.2.2 Attention Decomposition

During inference, we first compute the outputs of the shared bases, $\{Y_{\text{time},i}\}_{i=1}^{M}$ and $\{Y_{\text{space},j}\}_{j=1}^{N}$. The full output of any original attention head is then approximated by a linear combination of these bases plus two lightweight residuals—a positional residual and a free residual:

$$\widehat{Y} \approx \underbrace{\sum_{i=1}^{M} \alpha_i \, Y_{\text{time},i} + \sum_{j=1}^{N} \beta_j \, Y_{\text{space},j}}_{Y_{\text{base}}} + R_{\text{pos}} + R_{\text{free}} , \tag{1}$$

where $\{\alpha_i\}$ and $\{\beta_j\}$ are dynamically computed scalar weights. R_{pos} focuses on restoring high-frequency/positional signals (e.g., RoPE-contributed components), while R_{free} compensates residual content not captured by the shared bases under an orthogonality constraint. This decomposition shifts the computational bottleneck from L independent full-attention operations to just $M + N$ globally shareable basis operations, while preserving fidelity via two low-rank residuals.

3.3 Dynamic Weighting via Anchor Direction Projection: Achieving Content Awareness

While the bases provide a powerful vocabulary of patterns, the system requires an intelligent mechanism to dynamically compose them based on context. To this end, we adopt an **Anchor Direction Projection** mechanism to achieve content-aware allocation of the basis weights $\{\alpha_i, \beta_j\}$.

1) Defining anchor and candidate vectors. For an attention head, its input feature $h \in \mathbb{R}^{N_{\text{tok}} \times D}$ is flattened to serve as the anchor direction. The base outputs $\{Y_{\text{time},i}, Y_{\text{space},j}\}$ are similarly flattened to candidate vectors $\{y_k\}_{k=1}^{M+N}$.

2) Projection scores. We measure alignment by cosine similarity between each y_k and the anchor h:

$$s_k \;=\; \frac{\langle y_k,\, h\rangle}{\|y_k\|\,\|h\|}\,, \qquad k = 1, \ldots, M{+}N. \tag{2}$$

3) Decoupled weights. To avoid restricting the mixture to a probability simplex, we *do not* normalize the weights. We use a smooth nonnegative gate with a global gain:

$$\omega_k \;:=\; \gamma\,\sigma(\tau\, s_k)\,, \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid, $\tau > 0$ is a learnable temperature (shared or per-head), and $\gamma \geq 0$ is a lightweight gain (a scalar parameter). Thus $\omega_k \geq 0$ and $\sum_k \omega_k$ is *not* constrained to 1. Finally, $\{\omega_k\}$ are mapped to $\{\alpha_1, \ldots, \alpha_M, \beta_1, \ldots, \beta_N\}$. This preserves directional selection from cosine scores while providing smooth magnitude control without a simplex constraint.

3.4 Multifunctional Residual Module and Refined Training Objectives

Motivation. Although the shared-basis factorization is highly efficient, it is inherently a lossy compression. A single residual tasked with correcting this approximation often struggles, as it must simultaneously restore high-frequency positional signals (e.g., those from RoPE) and backfill low-frequency content not captured by the bases. This can lead to conflicting training signals and model degeneracy. We therefore disentangle the correction pathways by splitting the residual into two components with distinct responsibilities.

Formulation. We decompose the full attention approximation $\widehat{Y}$ into three parts: the shared basis combination (Y_{base}), a positional residual (R_{pos}), and a free residual (R_{free}). To maintain efficiency, both residuals are parameterized as low-rank tensors ($R_{\text{pos}} = \gamma_{\text{pos}}\, uv^\top$, $R_{\text{free}} = \gamma_{\text{free}}\, \tilde{u}\tilde{v}^\top$) predicted by a lightweight micro-network. This design achieves a clear "separation of concerns": R_{pos} is dedicated to restoring position-sensitive details, while R_{free} compensates for the remaining content error.

$$\widehat{Y} \;=\; Y_{\text{base}} \;+\; R_{\text{pos}} \;+\; R_{\text{free}} \tag{4}$$

Training Objective. To ensure each component learns its designated role without conflict, we assign them disjoint supervision targets. The training objective is a composite loss, with its components defined as follows:

$$\mathcal{L}_{\text{base}} := \left\| Y_{\text{full}}^{\text{no_pos}} - Y_{\text{base}} \right\|_F^2\,, \tag{5}$$

$$\mathcal{L}_{\text{pos}} := \| R_{\text{pos}} - \Delta_{\text{pos}} \|_F^2\,, \tag{6}$$

$$\mathcal{L}_{\text{free}} := \left\| \left(Y_{\text{full}} - Y_{\text{base}} - R_{\text{pos}}\right) - R_{\text{free}} \right\|_F^2\,, \tag{7}$$

$$\mathcal{L}_{\text{reg}} := \| P_{\mathcal{B}}(R_{\text{free}}) \|_F^2\,. \tag{8}$$

Here, $\mathcal{L}_{\text{base}}$ pushes the basis combination to fit the content signal; $\mathcal{L}_{\text{pos}}$ dedicates R_{pos} to exclusively match the pure positional signal ($\Delta_{\text{pos}} = Y_{\text{full}} - Y_{\text{full}}^{\text{no_pos}}$); $\mathcal{L}_{\text{free}}$ absorbs the remaining error; and $\mathcal{L}_{\text{reg}}$ uses a projection $P_{\mathcal{B}}$ onto the basis subspace to discourage R_{free} from relearning the bases. The total loss is a weighted sum of these components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda_{\text{pos}}\,\mathcal{L}_{\text{pos}} + \lambda_{\text{free}}\,\mathcal{L}_{\text{free}} + \lambda_{\text{reg}}\,\mathcal{L}_{\text{reg}}. \tag{9}$$

Complexity Analysis. For an H-head attention layer with sequence length N and head width d, standard MHA costs $\mathcal{O}(H\,N^2 d)$ per layer. Our DAD replaces H per-head attentions by $(M{+}N)$ shared bases:

$$\underbrace{\mathcal{O}\big((M{+}N)\,N^2 d\big)}_{\text{shared bases}} \;+\; \underbrace{\mathcal{O}(H\,d\,r)}_{\text{low-rank residuals}} \;+\; \underbrace{\mathcal{O}\big(H(M{+}N)d\big)}_{\text{light weighting}}.$$

Since $r \ll N$ and $(M{+}N) \ll H$, the dominant term drops from $\mathcal{O}(H\,N^2 d)$ to $\mathcal{O}((M{+}N)\,N^2 d)$, yielding an ideal speedup of $\approx H/(M{+}N)$ (up to small linear terms). In our settings ($H = 40$, $M{+}N = 8$, $r \leq 2$), this matches the observed $\sim 1.8\times$ end-to-end gain,as the overall speedup is limited by other unaccelerated components in the DiT block, such as the Feed-Forward Networks.

4 EXPERIMENTS

4.1 SETUPS

Models. We benchmark our method on two popular diffusion models: HunyuanVideo (Kong et al., 2024), and Wan2.1 (Wan et al., 2025), which 13, 1.3, and 14 billion parameters, respectively. HunyuanVideo can generate up to a 5-second video with 720p resolution and 125 frames. Wan2.1-14B can generate up to a 5-second video with 720p and 81 frames.

Metrics. Our evaluation covers both quality and efficiency. For quality, we compare against a full attention model using PSNR, LPIPS, and SSIM, while VBench (Huang et al., 2024) assesses background consistency and motion smoothness. For efficiency, we measure the sparse attention cost with density and the overall generation cost with total FLOPs.

Baselines. We compare against state-of-the-art sparse attention algorithms including static method Sparse VideoGen (SVG)(Xi et al., 2025), PA(Chen et al., 2025),dynamic methods SpargeAttention(Zhang et al., 2025a) and STA(Zhang et al., 2025b).

Table 1: Quantitative results at the default video length. Under the same computation budget, our method consistently outperforms STA, SparseAttn, and PA on VBench, PSNR, SSIM, and LPIPS, matches the fidelity of SVG, and achieves up to 1.96× speedup on HunyuanVideo and 1.82× on Wan2.1-1.3B with a single NVIDIA A100 GPU.

Model	Method	VBench (↑)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	TFLOPs	Latency (s)	Speedup
Hunyuan Video (50 frames)	Original	82.26	–	–	–	612	1724	–
	STA (FA2)	82.33	26.7	0.866	**0.167**	360	1120	1.54×
	SpargeAttn	82.31	26.5	0.862	0.175	350	1090	1.58×
	PA	81.90	22.1	0.764	0.256	355	1290	1.34×
	SVG	82.40	27.3	0.896	0.174	345	1150	1.50×
	Ours	**82.59**	**28.7**	**0.875**	0.214	**312**	**907**	**1.96×**
Wan2.1-1.3B-T2V (81 frames)	Original	81.80	–	–	–	560	469	–
	STA (FA2)	82.10	22.9	0.830	0.171	322	306	1.53×
	SpargeAttn	82.00	22.7	0.828	0.173	321	308	1.52×
	PA	81.50	22.4	0.790	0.176	324	287	1.63×
	SVG	81.90	23.2	0.812	0.202	324	293	1.60×
	Ours	**82.20**	**23.9**	**0.825**	**0.163**	**291**	**257**	**1.82×**
Wan2.1-1.3B-I2V (81 frames)	Original	82.36	–	–	–	880	482	–
	STA(FA2)	82.30	27.0	0.885	0.205	560	321	1.50×
	SpargeAttn	82.34	26.7	0.883	0.207	342	926	1.41×
	PA	81.90	26.2	0.870	0.220	565	295	1.63×
	SVG	82.40	27.4	0.890	0.210	570	303	1.59×
	Ours	**82.85**	**27.8**	**0.895**	**0.185**	**346**	**273**	**1.76×**
Wan2.1-14B-T2V (50 frames)	Original	82.36	–	–	–	1410	2100	–
	STA(FA2)	82.28	26.8	0.882	0.208	900	1458	1.44×
	SpargeAttn	82.30	26.6	0.880	0.210	905	1544	1.36×
	PA	81.85	26.0	0.868	0.224	905	1750	1.20×
	SVG	82.38	27.1	0.888	0.212	910	1373	1.53×
	Ours	**82.80**	**27.5**	**0.892**	**0.188**	**837**	**1105**	**1.85×**

Implementations. We perform benchmarks on an NVIDIA A100 GPU (40GB VRAM) with CUDA 12.4. Following previous work(Zhao et al., 2024b; Li et al., 2024b), all methods skip sparse attention for the first 20% of the denoising steps, as this stage is critical for generation quality.

4.2 QUALITY EVALUATION

To quantitatively assess the generation quality, we evaluate the quality of videos generated by DAD, when compared to baselines and report the results in Table 1 and Figure 5. DAD consistently outperforms all baseline methods in terms of PSNR, SSIM, and LPIPS, while still maintaining the

Table 2: Ablations on Wan 2.1 (1.3B) at default video length (simulated). We report VBench, Latency and Speedup.

Variant	VBench ↑	Latency (s)	Speedup
	Residual complementarity		
w/o residual R	81.70	273	1.72×
only low-rank R (no projection)	81.65	270	1.74×
projection + R (ours)	**82.20**	**257**	**1.82×**
	Scoring / projection		
energy / ℓ_2-norm scoring	81.90	272	1.73×
anchor-direction projection (ours)	**82.05**	**263**	**1.78×**
	Temperature τ sharing & robustness		
shared τ	**82.12**	**261**	**1.80×**
per-head τ	82.10	266	1.76×
shared-τ sweep: 0.5/1/2/4	82.05/82.12/82.20/82.08	270/263/257/266	1.74–1.82×

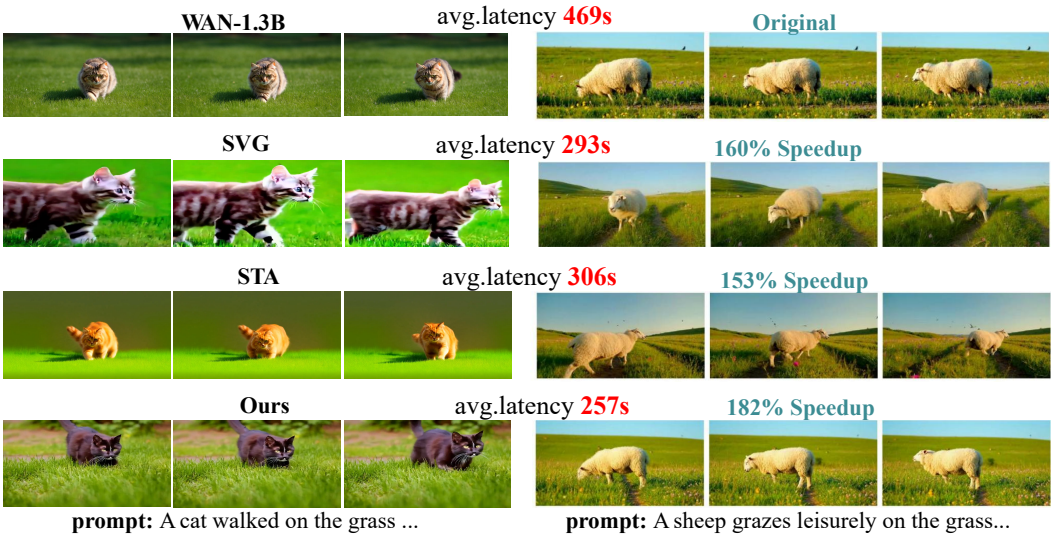


Figure 5: Qualitative comparison of sparse attention methods: SVG, STA, and Ours.

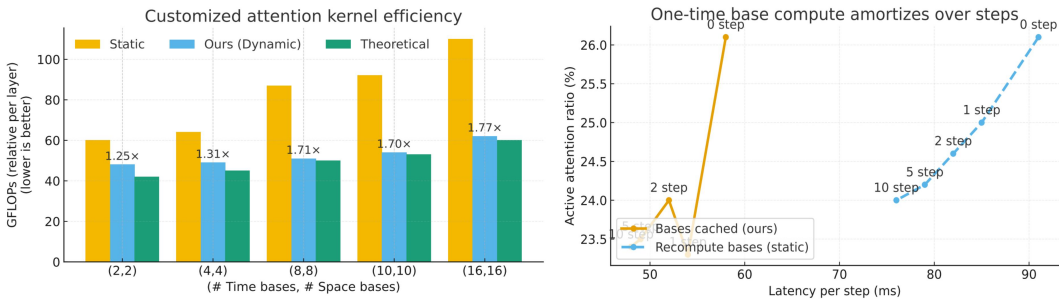
highest speedup. Specifically, DAD achieves an average PSNR of 28.3 on HunyuanVideo and 23.9 on Wan2.1, demonstrating its effectiveness on generating highly consistent and smooth videos while providing a 1.96× speedup on HunyuanVideo and 1.82× on Wan2.1-1.3B-T2V.

4.3 Ablation Study

We perform ablation studies on Wan 2.1 (1.3B) to assess the contributions of various components to the model's performance, as shown in Table 2. Our method demonstrates that incorporating residual complementarity with projection improves VBench performance (82.20) and reduces latency (257s), achieving a 1.82× speedup. We further observe that using anchor-direction projection boosts speedup and reduces latency compared to the energy/ℓ2-norm scoring approach. Additionally, sharing temperature τ across heads provides a slight improvement in both performance and speedup. These results confirm the effectiveness of our proposed design choices in enhancing both efficiency and generation quality.

4.4 Efficiency of Dynamic Routing and Amortization

We perform a detailed efficiency analysis to verify the core components of our proposed method (Figure 6). As shown in Figure 6 (a), our DAD framework featuring dynamic routing (Ours) is



(a) (b)

Figure 6: **Efficiency of shared spatio-temporal bases.** (a) Computing once and reusing bases with dynamic routing (*Ours*) reduces GFLOPs and remains close to the theoretical lower bound across different (T,S). (b) Per-step latency vs. active-attention ratio over diffusion steps: caching amortizes the one-time base compute, yielding consistently lower latency and a smaller active ratio than recomputing at every step.

compared against a static-weighted baseline (Static). The results indicate that dynamic routing consistently reduces per-layer computation (GFLOPs), with performance closely approaching the theoretical lower bound (Theoretical), which proves our method's low overhead. Furthermore, Figure 6 (b) highlights the benefits of amortization across the diffusion process. A one-time computation and caching of the shared bases, despite a higher initial latency at step 0, allows our method to achieve substantially lower per-step latency and a smaller active attention ratio in subsequent steps, clearly outperforming the static approach that recomputes the bases at every step.

4.5 Basis Trade-offs

We analyze the trade-off between the number of shared bases ($T + S$) and the resulting generation quality and inference speed, as shown in Figure 7. The results reveal that both PSNR and SSIM metrics improve rapidly as we increase the number of bases from two to approximately eight. Conversely, the inference speedup shows a corresponding decrease, dropping from 2.4× to 1.8× in the same range. Since further increasing the number of bases beyond eight yields diminishing returns in quality, we identify $T + S \approx 8$ as the optimal operating point that balances efficiency and fidelity.

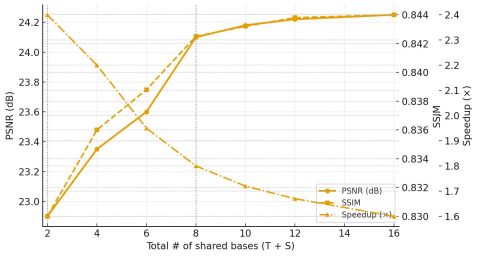


Figure 7: Trade-off between the number of shared bases and quality/speed (WAN-2.1-1.3B, simulated). As the total bases $T+S$ increase, PSNR/SSIM rise quickly with few bases and plateau around $T+S \approx 8$; further growth yields negligible gains. Meanwhile, the speedup drops from about 2.4× to 1.6×, indicating $T+S \approx 8$ is a good operating point balancing quality and efficiency.

5 Conclusion & Limitation

In this paper, we proposed the DAD framework, a novel and efficient inference acceleration method for DiT-based video generation. Our core insight is that the functionality of numerous attention heads can be reconstructed from a compact, shared set of spatio-temporal bases. By dynamically combining these bases using a content-aware anchor projection mechanism and compensating for approximation errors with a multi-functional residual module, DAD dramatically reduces computational redundancy. Comprehensive evaluations demonstrate that our framework achieves a superior trade-off between generation quality and efficiency, delivering significant speedups while maintaining high fidelity. The primary limitation of this work is that the basis selection process is performed offline; future work could explore methods for learning or adapting these bases online, potentially offering greater flexibility for diverse, unseen data distributions.

REFERENCES

Karim Ahmed, Nitish Shirish Keskar, and Richard Socher. Weighted transformer network for machine translation. *arXiv preprint arXiv:1711.02132*, 2017.

Lida Chen, Dong Xu, Chenxin An, Xintao Wang, Yikai Zhang, Jiangjie Chen, Zujie Liang, Feng Wei, Jiaqing Liang, Yanghua Xiao, et al. Powerattention: Exponentially scaling of receptive fields for effective sparse attention. *arXiv preprint arXiv:2503.03588*, 2025.

Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. Multi-head attention: Collaborate instead of concatenate. *arXiv preprint arXiv:2006.16362*, 2020.

Dongchen Han, Ziyi Wang, Zhuofan Xia, Yizeng Han, Yifan Pu, Chunjiang Ge, Jun Song, Shiji Song, Bo Zheng, and Gao Huang. Demystify mamba in vision: A linear attention perspective. *Advances in neural information processing systems*, 37:127181–127203, 2024.

Jingcheng Hu, Houyi Li, Yinmin Zhang, Zili Wang, Shuigeng Zhou, Xiangyu Zhang, Heung-Yeung Shum, and Daxin Jiang. Multi-matrix factorization attention. *arXiv preprint arXiv:2412.19255*, 2024.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36:24892–24928, 2023.

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

Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, and Xun Zhou. Flexprefill: A context-aware sparse attention mechanism for efficient long-sequence inference. *arXiv preprint arXiv:2502.20766*, 2025.

Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *Advances in neural information processing systems*, 37:75692–75726, 2024a.

Jian Li, Baosong Yang, Zi-Yi Dou, Xing Wang, Michael R Lyu, and Zhaopeng Tu. Information aggregation for multi-head attention with routing-by-agreement. *arXiv preprint arXiv:1904.03100*, 2019.

Muyang Li, Tianle Cai, Jiaxin Cao, Qinsheng Zhang, Han Cai, Junjie Bai, Yangqing Jia, Kai Li, and Song Han. Distrifusion: Distributed parallel inference for high-resolution diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7183–7193, 2024b.

Xingyang Li, Muyang Li, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, et al. Radial attention: $o(n\backslash$log n) $sparseattentionwithenergydecayforlongvideogeneration.$*arXiv preprint arXiv:2506.19852*, $2025a.$

Yucheng Li, Huiqiang Jiang, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Amir H Abdi, Dongsheng Li, Jianfeng Gao, Yuqing Yang, et al. Mminference: Accelerating prefilling for long-context vlms via modality-aware permutation sparse attention. *arXiv preprint arXiv:2504.16083*, 2025b.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. 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.

Xiaoyu Liu, Jiahao Su, and Furong Huang. Tuformer: Data-driven design of transformers for improved generalization or efficiency. In *The Tenth International Conference on Learning Representations (ICLR 2022)*, 2022.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15762–15772, 2024.

Tan Nguyen, Tam Nguyen, Hai Do, Khai Nguyen, Vishwanath Saragadam, Minh Pham, Khuong Duy Nguyen, Nhat Ho, and Stanley Osher. Improving transformer with an admixture of attention heads. *Advances in neural information processing systems*, 35:27937–27952, 2022.

Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3531–3539, 2021.

Wenhao Sun, Rong-Cheng Tu, Yifu Ding, Zhao Jin, Jingyi Liao, Shunyu Liu, and Dacheng Tao. Vorta: Efficient video diffusion via routing sparse attention. *arXiv preprint arXiv:2505.18809*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, et al. Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. *arXiv preprint arXiv:2502.01776*, 2025.

Yifei Xia, Suhan Ling, Fangcheng Fu, Yujie Wang, Huixia Li, Xuefeng Xiao, and Bin Cui. Training-free and adaptive sparse attention for efficient long video generation. *arXiv preprint arXiv:2502.21079*, 2025.

Ruyi Xu, Guangxuan Xiao, Haofeng Huang, Junxian Guo, and Song Han. Xattention: Block sparse attention with antidiagonal scoring. *arXiv preprint arXiv:2503.16428*, 2025.

Shuo Yang, Haocheng Xi, Yilong Zhao, Muyang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chenfeng Xu, Kelly Peng, et al. Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation. *arXiv preprint arXiv:2505.18875*, 2025.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22963–22974, 2025.

Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. Spargeattn: Accurate sparse attention accelerating any model inference. *arXiv preprint arXiv:2502.18137*, 2025a.

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

Xiaofeng Zhang, Yikang Shen, Zeyu Huang, Jie Zhou, Wenge Rong, and Zhang Xiong. Mixture of attention heads: Selecting attention heads per token. *arXiv preprint arXiv:2210.05144*, 2022.

Tianchen Zhao, Tongcheng Fang, Haofeng Huang, Enshu Liu, Rui Wan, Widyadewi Soedarmadji, Shiyao Li, Zinan Lin, Guohao Dai, Shengen Yan, et al. Vidit-q: Efficient and accurate quantization of diffusion transformers for image and video generation. *arXiv preprint arXiv:2406.02540*, 2024a.

Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. Real-time video generation with pyramid attention broadcast. *arXiv preprint arXiv:2408.12588*, 2024b.

Shihao Zhou, Dayu Li, Jinshan Pan, Juncheng Zhou, Jinglei Shi, and Jufeng Yang. Devil is in the uniformity: Exploring diverse learners within transformer for image restoration. *arXiv preprint arXiv:2503.20174*, 2025.

A APPENDIX

A.1 THE USE OF LLM

We utilized Large Language Model (LLM) to assist in the preparation of this manuscript. The primary role of the LLM was to improve the clarity, conciseness, and grammatical accuracy of the text. Its application was strictly limited to refining and rephrasing sentences and paragraphs that were already written. The LLM was not used for generating original ideas, writing entire sections, or producing any of the code presented in this work. All intellectual contributions, including the core concepts and experimental results, are entirely our own.