ROPESLR: 3D ROPE-DRIVEN SPARSE-LOWRANK ATTENTION FOR EFFICIENT DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion Transformers (DiTs) have become dominant for high-fidelity video generation, yet their $\mathcal{O}(L^2)$ attention cost bottlenecks long-sequence synthesis. Recent sparse-linear hybrids reduce this cost but degrade at extreme sparsity due to what we call the *RoPE dilemma*: positive feature maps enabling linear attention do not preserve 3D Rotary Position Embeddings (RoPE) orthogonality, losing explicit relative spatiotemporal offsets in global aggregation. We propose **RoPeSLR**, a 3D RoPE-aware sparse-low-rank framework. Through structural analysis, we observe that DiT attention matrices empirically exhibit decomposition into high-energy sparse spikes and smooth backgrounds with slowly growing rank. Motivated by this observation, we design an energy-driven block-sparse branch for dominant interactions and a Fourier low-rank compensator preserving RoPE structure via rotated Fourier features aggregating global context in $\mathcal{O}(Lrd_{\text{head}})$. At 90% sparsity, RoPeSLR maintains strong VBench quality on Wan2.1-1.3B and 14B while reducing DiT denoising latency by $1.7\times$ and $1.8\times$, respectively.

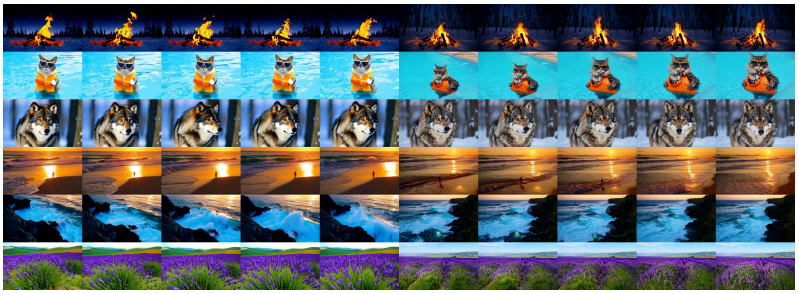

Figure 1: Video generation: full attention vs. RoPeSLR at 90% sparsity on Wan2.1-1.3B.

1 INTRODUCTION

Diffusion Transformers (DiTs) Peebles & Xie (2023) have become the dominant architecture for high-fidelity video generation. However, processing the full spatiotemporal token sequence introduces a severe computational bottleneck, because the attention mechanism costs $\mathcal{O}(L^2)$ and scales quadratically with the sequence length L.

To mitigate this bottleneck, sparse attention mechanisms have been widely adopted to restrict queries to localized neighborhoods Xi et al. (2025); Chen et al. (2025a); Yang et al. (2025a). While efficient, pushing sparsity to extreme levels masks out global receptive fields and truncates long-range dependencies. In practice this often leads to semantic misalignment and to loss of fine-grained detail, which places a ceiling on the usable sparsity.

Recent work attempts to recover lost global context by fusing sparse and linear attention Zhang et al. (2025a). Though effective at moderate sparsity, these hybrids face two limitations at extreme sparsity: (1) **The RoPE dilemma:** to maintain matrix associativity, they rely on positive feature maps that *fail to preserve* 3D RoPE's orthogonal relative-position structure. Consequently, global context aggregation loses explicit relative spatiotemporal offsets, weakening distance awareness. (2) **Limited structural justification:** these heuristic designs offer little explanation of *why* a low-complexity term can approximate non-linear softmax attention without a severe distribution shift.

This brings us to a pivotal question:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

> **Question.** *Is there an efficient attention design that stays close to full attention at extreme sparsity, is informed by structural analysis of DiT attention patterns rather than pure heuristics, and improves the quality–efficiency trade-off in downstream video generation?*

To answer this, we propose **RoPeSLR**, a 3D RoPE-aware sparse-low-rank attention framework motivated by structural analysis (**Proposition 1**). A bounded-norm existence result supports sparse-plus-low-rank reconstruction in expected entry-wise error, while measurements on pre-trained DiTs show that interaction magnitudes concentrate on leading frequency pairs. Together, these findings suggest separating high-energy semantic spikes from a smooth low-rank background.

RoPeSLR instantiates this principle: an energy-driven block-sparse branch computes high-energy spikes explicitly, while a Fourier low-rank compensator preserves RoPE structure via rotated Fourier features, aggregating global context in $\mathcal{O}(Lrd_{\text{head}})$. This design (1) mirrors the structural observations, (2) retains RoPE-induced relative-position dependence in its features through embedded rotations, and (3) preserves non-linear expressivity via positive kernels.

Our main contributions are:

- **Structural analysis.** We empirically identify sparse-plus-low-rank structure in DiT attention maps with 3D RoPE, providing asymptotic scaling arguments and empirical validation.
- **Architecture.** RoPeSLR decouples responsibilities: energy-driven block-sparse branch for speedup, Fourier low-rank compensator for global context with RoPE preservation.
- **Results.** Strong VBench quality on Wan2.1-1.3B/14B at 90% sparsity with substantial latency reduction.

Related Work is in Appendix A.

2 PRELIMINARY

Notation and Attention Paradigms. Let $X \in \mathbb{R}^{L\times d_{\text{model}}}$ denote a video sequence flattened from a spatiotemporal grid $p = (t, h, w) \in [T] \times [H] \times [W]$. For each attention head, the input is projected into $Q, K, V \in \mathbb{R}^{L\times d_{\text{head}}}$. Token-specific representations at position p are denoted as $q_p, k_p, v_p \in \mathbb{R}^{d_{\text{head}}}$.

Standard attention computes $O = \text{softmax}(QK^\top/\sqrt{d_{\text{head}}})V$, which costs $\mathcal{O}(L^2)$ and becomes the dominant cost in video generation. Let $A_{p,q}$ denote the post-softmax attention weight from token p to token q. For each head h, we write q_p^h, k_p^h, v_p^h for the per-head query, key, and value vectors, and use $s_{p,q}^h$ and $A_{p,q}^h$ for the corresponding pre-softmax logit and attention weight. Throughout the main text, $R(\cdot)$ *with* an argument always denotes a rotation matrix, whereas R *without* an argument denotes a rank.

Block-sparse attention circumvents the quadratic cost by restricting queries to a localized neighborhood $\mathcal{N}(p)$:

$$O_p = \sum_{q\in\mathcal{N}(p)} \frac{\exp\left(q_p^\top k_q/\sqrt{d_{\text{head}}}\right)}{\sum_{r\in\mathcal{N}(p)} \exp\left(q_p^\top k_r/\sqrt{d_{\text{head}}}\right)} v_q. \tag{1}$$

3D RoPE and Shift-Invariance. 3D RoPE imparts geometric priors via orthogonal rotation. The head dimension is partitioned across axes ($d_{\text{head}} = d_t + d_h + d_w$), with rotational frequencies decaying exponentially in the pair index m:

$$\theta_m^k = 10000^{-2(m-1)/d_k}, \quad k \in \{t, h, w\}, \quad m \in [1, d_k/2]. \tag{2}$$

Thus small m corresponds to a large rotational frequency and large m to a small one. The composite rotation matrix $R(p) \in \mathbb{R}^{d_{\text{head}}\times d_{\text{head}}}$ is block-diagonal, and the pre-softmax logit is shift-invariant, which is what makes its value depend only on the relative position:

$$s_{p,q} = \frac{1}{\sqrt{d_{\text{head}}}}\langle R(p)q_p, R(q)k_q\rangle = \frac{1}{\sqrt{d_{\text{head}}}}\langle q_p, R(q-p)k_q\rangle. \tag{3}$$

Linear Attention and The RoPE Dilemma. Linear attention achieves $\mathcal{O}(L)$ complexity by linearizing the exponential kernel with a feature map $\phi(\cdot) : \mathbb{R}^{d_{\text{head}}} \to \mathbb{R}^r$, exploiting matrix associativity:

$$O_p = \phi(q_p)^\top \underbrace{\sum_{q=1}^{L} \phi(k_q) v_q^\top}_{\text{Global Context } C}. \tag{4}$$

The Dilemma. The global aggregation of the context C removes the explicit dependence on the relative offset $(p - q)$. Because the positive feature maps used for matrix associativity generally do not commute with the orthogonal rotation ($\phi(R(p)q_p)^\top \phi(R(q)k_q) \neq \phi(q_p)^\top \phi(k_q) \cdot f(p - q)$), the RoPE structure cannot be factored out of the summation. This forces a trade-off between $\mathcal{O}(L)$ efficiency and relative positional cues, and motivates a compensator that factorizes RoPE into *explicit* rotated Fourier bases instead of relying on a generic feature map.

3 METHOD

RoPeSLR is designed to exploit the sparse-plus-low-rank structure of attention maps in Diffusion Transformers (DiTs).

3.1 STRUCTURAL ANALYSIS FOR DESIGN

Our architecture is motivated by complementary structural results for the post-softmax attention map A in 3D RoPE-equipped DiTs. Appendix B proves a sparse-plus-low-rank reconstruction guarantee under bounded query/key norms and separately analyzes Fourier truncation under the empirically supported spectral prior in Assumption 2.

Proposition 1 (Design-motivating structural analysis, informal)**.** *Under Assumption 1, choose* $\tau = cL^{-1/2}$ *and* $\mathcal{E} = \tau/2$. *Let* $\Omega_\tau = \{(p, q) : A(p, q) > \tau\}$. *There exists a randomized reconstruction* $\hat{A}_{final} = \tilde{A}_{sparse} + \hat{A}_{lowrank}$ *such that:*

1. ***Sub-quadratic sparsity.*** *The residual sparse part corrects the reconstruction exactly on* Ω_τ *and has* $\text{NNZ}(\tilde{A}_{sparse}) = \mathcal{O}(L^{3/2})$.
2. ***Sequence-length-independent rank.*** *For fixed query/key norm bound and fixed ratio* $\tau/\mathcal{E}$, $\text{rank}(\hat{A}_{lowrank})$ *is bounded independently of* L.
3. ***Expected entry-wise error.*** *Over the positive-random-feature draw,*

$$\max_{p,q} \left(\mathbb{E} |\hat{A}_{final}(p, q) - A(p, q)|^2 \right)^{1/2} \leq \mathcal{E} = \mathcal{O}(L^{-1/2}).$$

The feature-rank constant depends exponentially on the query/key norm bound, so this is an existence result rather than a practical rank estimate.

Role of the analysis.. The analysis serves three purposes. First, it identifies a sparse-plus-low-rank organization in DiT attention: high-energy entries are few, while the residual background is empirically smooth. Second, Assumption 2 and the Fourier expansion explain why truncating frequency components is a plausible compact design, without attributing the observed frequency-index decay to prior work. Third, the proposition gives a rigorous idealized reconstruction guarantee using the finite second moment of positive random features. RoPeSLR realizes this blueprint with practical components for training and inference; Table 1 summarizes the correspondence.

Analysis–architecture correspondence.. The analysis characterizes an *idealized* reconstruction; the deployed architecture relaxes it in three explicit ways. (i) *Entry-wise thresholding becomes block-wise selection* for hardware efficiency; quality is preserved across sparsity levels (Table 4). (ii) *FAVOR+ random features become a learnable positive map*; the theoretical random-feature guarantee is in expectation, while the learned map is evaluated empirically. Its branch inputs $(z_p^h)^\top y_q^h$ remain exactly relative-offset dependent (Figure 4). (iii) *Exact normalization with complementary supports becomes independent normalization with a learned gate*; the gate learns content-dependent allocation between the two branches (Figure 10). The proposition is therefore a design blueprint rather than a certificate of the deployed system.

Table 1: Correspondence between the structural analysis and the deployed RoPeSLR architecture.

Structural analysis	Design principle	RoPeSLR instantiation	Evidence
Sparse spike set Ω_τ	High-energy interactions should be computed explicitly	Energy-driven block-sparse branch computes exact softmax attention on retained blocks	Table 3, Table 4
Smooth background A_{bg}	Low-energy background can be summarized by a compact global state	Fourier low-rank branch aggregates u^h and C^h with rank r	Figure 11
3D RoPE Fourier expansion	Global aggregation should remain relative-position aware	Rotated Fourier features z_p^h, y_q^h and positive kernel approximation	Figure 4
Idealized reconstruction	Practical training requires stable fusion	Learned gate and branch-wise normalization	Table 3

From the analysis to the architecture.. The constructive part of the analysis suggests a two-branch design: high-energy entries are handled explicitly, while the smooth background is approximated by truncating the spectrally dominant RoPE Fourier series and linearizing the resulting exponential kernel with positive features. RoPeSLR follows this blueprint and makes three practical choices for trainable video generation: (i) block-wise energy selection replaces entry-wise thresholding to obtain hardware-friendly sparse kernels; (ii) a learnable positive projection replaces fixed random features to improve optimization stability; and (iii) a gated fusion combines the two branches under independent normalization. These choices preserve the same computational division of labor while adapting the idealized construction to large-scale training and inference. We instantiate the two components in Sections 3.2 and 3.3.

Mathematical Anatomy of 3D RoPE.. By orthogonal shift-invariance ($R(p)^\top R(q) = R(q-p)$), the pre-softmax logit admits an *exact* Fourier series expansion over the relative spatiotemporal offset Δ, with $\Delta_k = q_k - p_k$ (Lemma 2):

$$s_{p,q} = \sum_{k\in\{t,h,w\}} \sum_{m=1}^{d_k/2} \left(a^k_{p,q,m} \cos(\theta^k_m \Delta_k) + b^k_{p,q,m} \sin(\theta^k_m \Delta_k) \right), \tag{5}$$

where we omit the constant $1/\sqrt{d_{\mathrm{head}}}$, and the coefficients $a^k_{p,q,m}, b^k_{p,q,m}$ depend on the pair (p,q) only through the corresponding 2-D frequency-subspace projections of q_p and k_q. Under this basis, our measurements on pre-trained DiTs indicate that the interaction magnitudes decay quickly with the pair index m (Appendix C.1, Figure 5). Assumption 2 formalizes this observation as a structural prior, which makes the smooth background amenable to a low-rank surrogate while the residual high-energy interactions are handled sparsely.

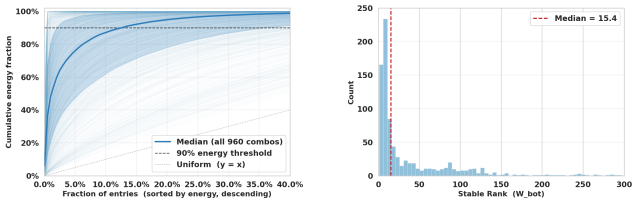

Figure 2: Sparse/low-rank profile of HunyuanVideo post-softmax attention across 960 layer–head combinations. Each row of the attention matrix is split by a 90% per-row cumulative-energy criterion, and the low-energy residual is analyzed. This per-row criterion is a practical proxy for the thresholding idea used in the structural analysis.

Comparison between DiTs and LLMs.. We also provide a structural comparison between DiTs and LLMs. First, in the truncation step of Lemma 3, the required Fourier index scales with $1/(-\log\rho)$; splitting the head dimension across three axes cubes the per-axis factor ρ, which reduces the truncation index by a constant factor under a uniform split. Second, the deterministic sparsity bound becomes practically useful when high-energy interactions are clustered; the spatiotemporal locality of video encourages such clustering, whereas causal language-model attention can scatter high-energy entries across the context. We report this comparison as supporting evidence rather than a formal separation; see Appendix B, Remark 2.

Empirical Validation. Attention maps of pre-trained DiTs exhibit the sparse-plus-low-rank profile across layers and denoising steps (Figure 2). As the sequence length L grows, the stable rank $\mathrm{sr}(M) = \|M\|_F^2/\|M\|_2^2$ of the low-energy residual grows sublinearly over the tested range (Figure 3). A comparison with Qwen2-7B-Instruct under a matched 90% per-row energy protocol further suggests that this decomposition is more pronounced in the DiTs we probe (Figure 9).

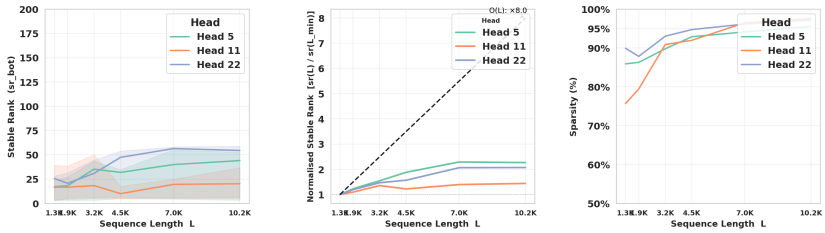

Figure 3: HunyuanVideo post-softmax matrices decomposed using a 90% per-row energy threshold. Across sequence lengths from 1.3K to 10.2K, the low-energy residual has bounded stable rank with sublinear empirical growth.

3.2 Fourier Low-Rank Compensator

Standard linear attention does not preserve the relative positional offsets carried by 3D RoPE, and therefore weakens distance awareness (Section 2). To recover this without reintroducing $\mathcal{O}(L^2)$ cost, we follow the constructive route suggested by the Fourier expansion of 3D RoPE (Lemma 2) and by the low-rank kernel linearization used in the analysis.

For each head h, let q_p^h, k_p^h, v_p^h denote the query, key, and value vectors. Let $\mathcal{I}^h$ denote the *retained set of leading (spectrally dominant) RoPE frequency pairs*, i.e. the pairs with small index m that carry most of the measured interaction magnitude, consistent with the truncation in Lemma 3; in practice $\mathcal{I}^h$ is taken to be the leading frequency pairs of each axis, following the spectral measurements in Appendix C.1. Let $q_{p,\mathcal{I}}^h$ and $k_{q,\mathcal{I}}^h$ be the query/key subvectors corresponding to $\mathcal{I}^h$, and let $R_{\mathcal{I}}^h(p)$ be the RoPE rotation restricted to these pairs. We define the absolute rotated Fourier features

$$z_p^h = R_{\mathcal{I}}^h(p)\, q_{p,\mathcal{I}}^h, \qquad y_q^h = R_{\mathcal{I}}^h(q)\, k_{q,\mathcal{I}}^h. \tag{6}$$

By orthogonality of the RoPE rotations,

$$(z_p^h)^\top y_q^h = s_{p,q}^{(M),h}, \tag{7}$$

where $s_{p,q}^{(M),h}$ is the truncated pre-softmax logit that retains only the frequency pairs in $\mathcal{I}^h$ (up to the constant $1/\sqrt{d_{\mathrm{head}}}$, which we fold into the subsequent feature map). Thus, although z_p^h and y_q^h are built from *absolute* coordinates, their inner product depends only on the relative offset $p-q$.

We then apply a head-specific positive feature map Φ^h with feature dimension r:

$$\phi_p^h = \Phi^h(z_p^h), \qquad \psi_q^h = \Phi^h(y_q^h), \qquad \phi_p^h, \psi_q^h \in \mathbb{R}_+^r, \tag{8}$$

where r denotes the practical feature dimension used by the low-rank branch; in our experiments, we select $r = 64$ empirically rather than treating it as a strict theoretical rank bound. The design target is

$$\kappa^h(p,q) = (\phi_p^h)^\top \psi_q^h \;\approx\; \exp\!\left(s_{p,q}^{(M),h}\right). \tag{9}$$

This yields an explicit low-rank surrogate for the background attention kernel. Equation 9 is the design target of the low-rank branch. When Φ^h is instantiated with FAVOR+ random features Choromanski et al. (2021), this target holds in expectation and the induced kernel remains shift-invariant in $p-q$. Our default implementation uses a lightweight learnable positive projection trained toward the same target; this choice improves optimization stability in large-scale post-training while relaxing exact kernel-level shift-invariance, and we verify that distance-dependent behavior is retained at the output level (Figure 4; activation choices are ablated in Appendix C.2.3).

Empirical Validation. At the output level, RoPeSLR reproduces the distance-dependent similarity pattern of full attention (Figure 4). Mechanistically, the Gram spectral analysis in **Appendix**

C.6 shows that the principal spatial modes of the low-rank branch are spatially smooth, i.e. concentrated at low *spatial* frequencies of the latent grid, which is consistent with its intended role as a background approximator. This refers to the smoothness of the branch *output* in space, and is distinct from the RoPE pair index m used to define $\mathcal{I}^h$.

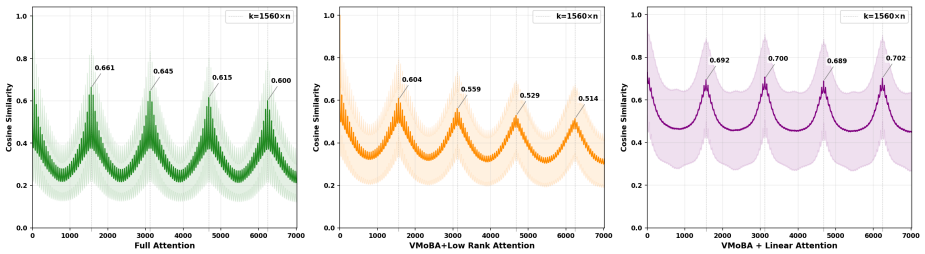

Figure 4: Cosine similarity between the per-token outputs of the three attention variants compared in this work, aggregated over token pairs and plotted against the relative token distance. The periodic pattern reflects the distance dependence induced by 3D RoPE.

3.3 Sparse-Low-Rank Attention

The analysis suggests a two-branch design: the sparse branch captures high-energy semantic spikes, while the low-rank branch summarizes the smooth global background. Both branches are head-specific, allowing different heads to specialize at different spatiotemporal scales.

Energy-driven sparse branch. The analysis identifies the spike set Ω_τ. We approximate it by an energy-thresholded block-sparse selection. For each head h and query token p, let $\mathcal{S}_p^h$ denote the set of retained keys. In our implementation, tokens are grouped into blocks, the block-level energy is estimated from pooled query/key representations, and only blocks whose estimated energy exceeds the threshold τ are evaluated, so that the full $L \times L$ attention matrix is never materialized. The sparse branch then computes attention exactly on the retained set:

$$O_{\mathrm{sp},p}^{(h)} = \sum_{q \in \mathcal{S}_p^h} \frac{\exp(s_{p,q}^h)}{\sum_{r \in \mathcal{S}_p^h} \exp(s_{p,r}^h)} v_q^h. \tag{10}$$

This branch allocates computation adaptively to high-energy interactions. The analysis suggests an adaptive budget that grows as $|\mathcal{S}_p^h| = \mathcal{O}(\sqrt{L})$ to obtain a sub-quadratic asymptotic regime. For stable inference on current video models, we instead use a fixed high sparsity ratio, which already yields a favorable quality–efficiency trade-off; the adaptive regime is an interesting direction for further scaling (Appendix D).

Global low-rank branch. Using the positive Fourier features from Section 3.2, we aggregate two sequence-level states:

$$u^h = \sum_{q=1}^{L} \psi_q^h, \qquad C^h = \sum_{q=1}^{L} \psi_q^h (v_q^h)^\top. \tag{11}$$

The low-rank output is a normalized readout of this global state:

$$O_{\mathrm{lr},p}^{(h)} = \frac{(\phi_p^h)^\top C^h}{(\phi_p^h)^\top u^h + \epsilon}, \tag{12}$$

where $\epsilon > 0$ prevents division by zero. Equivalently, let $\Phi^h(Z^h), \Phi^h(Y^h) \in \mathbb{R}^{L \times r}$ be the matrices whose rows are $(\phi_p^h)^\top$ and $(\psi_q^h)^\top$, respectively. Then this branch implements

$$\hat{A}_{\mathrm{lr}}^{(h)} = D^h \Phi^h(Z^h) \Phi^h(Y^h)^\top, \tag{13}$$

where D^h is the diagonal row-normalization matrix with entries

$$D_{pp}^h = \frac{1}{(\phi_p^h)^\top u^h + \epsilon},$$

so that $\mathrm{rank}(\hat{A}^{(h)}_{\mathrm{lr}}) \le r$. No $L \times L$ matrix is materialized: the state C^h has size $r \times d_{\mathrm{head}}$, giving $\mathcal{O}(Lrd_{\mathrm{head}})$ computation. The branch is global because u^h and C^h summarize all tokens, and its features retain an explicit dependence on the relative offset because the pre-map inner product $(z^h_p)^\top y^h_q$ is a function of $p-q$ only. The normalization uses the surrogate kernel of the low-rank branch, which makes the branch trainable without materializing the full partition function. This design choice keeps the low-rank branch efficient and stable, while the Stage-I alignment objective ensures that its output remains compatible with the sparse branch.

3.4 Gated Fusion

Because the two branches are normalized independently, fusing localized sparse features with the global low-rank context can introduce a scale mismatch. We therefore use a lightweight gated fusion. A token-wise scalar gate, shared across heads, is computed directly from the input tokens:

$$g = \sigma(XW_g + b_g), \qquad W_g \in \mathbb{R}^{d_{\text{model}} \times 1}, \quad b_g \in \mathbb{R}, \tag{14}$$

where σ is the sigmoid function. The two branch outputs are then fused and stabilized by RMSNorm, applied per head over the d_{head} channels:

$$O^{(h)} = \mathrm{RMSNorm}\left(O^{(h)}_{\mathrm{sp}} + g \odot O^{(h)}_{\mathrm{lr}}\right). \tag{15}$$

The gate controls how much global low-rank context is injected into the sparse representation, while RMSNorm keeps the fused representation numerically stable. Note that the gate operates directly on the input tokens without additional positional embeddings, as the low-rank branch already encodes relative spatiotemporal structure in its features via the internal RoPE-derived rotations. In practice, initializing the newly added parameters so that the fused output starts close to the sparse-branch output makes Stage-I alignment more stable; we adopt this strategy throughout our experiments. Compared with an ideal residual correction, the learned gate provides a stable and trainable way to allocate capacity between exact local attention and global low-rank context. This design keeps the sparse branch dominant at initialization while allowing the low-rank branch to restore global context during post-training.

3.5 Post-training Strategy

RoPeSLR is integrated into pre-trained DiTs via a lightweight two-stage post-training pipeline.

Stage I: Component Alignment. We freeze the pre-trained backbone and optimize only the newly introduced parameters

$$\Theta_{\text{new}} = \{\Theta^{(h)}_{\Phi}, \alpha_{\mathrm{pe}}, W_g, b_g, \Theta_{\text{norm}}\},$$

where $\Theta^{(h)}_{\Phi}$ denotes the head-specific positive Fourier feature projections, α_{pe} is a learnable scalar that rescales the rotated Fourier features before the positive map, and Θ_{norm} the learnable RMSNorm parameters. To warm-start the decomposition without perturbing the pre-trained representation, we minimize the mean squared error between our sparse-low-rank output O and the exact full attention output O_{full}:

$$\mathcal{L}_{\text{align}} = \frac{1}{L \cdot d_{\text{model}}} \|O - O_{\text{full}}\|^2_F. \tag{16}$$

Stage II: Full Fine-tuning. After alignment, we unfreeze the backbone and jointly optimize all parameters with the task-specific diffusion objective $\mathcal{L}_{\text{task}}$. This stage recovers end-to-end generation quality while preserving the efficiency of RoPeSLR.

Output Alignment Analysis. Ablations on the training schedule (Appendix C.2.4) further indicate that the two-stage pipeline is robust to the Stage-I duration in the range we tested.

4 Experiments

4.1 Experimental Settings

We evaluate RoPeSLR on Wan2.1-T2V-1.3B and Wan2.1-T2V-14B pre-trained DiT models at 90% sparsity. Wan2.1-1.3B is evaluated at 480p, while Wan2.1-14B is evaluated at 720p with 81 frames.

Training follows a two-stage post-training strategy: Stage-I aligns low-rank outputs with pre-trained features, Stage-II fine-tunes the complete system. Complete settings are detailed in the Appendix.

4.2 Results and Analysis

We evaluate RoPeSLR on Wan2.1-1.3B and Wan2.1-14B Wan et al. (2025) to assess cross-scale generality. For fine-tuning, we construct model-specific curated video–text subsets from OpenSora Zheng et al. (2024) and OpenVid-1M Nan et al. (2024): Wan2.1-1.3B uses an OpenSora subset, while Wan2.1-14B uses an OpenVid-1M subset. The subsets contain diverse, high-quality clips and are preprocessed to match each model's resolution and frame configuration.[1]

Baselines. We compare RoPeSLR with recent sparse and sparse-linear attention approaches:

- **SLA** Zhang et al. (2025a): a fine-tunable sparse-linear attention framework that fuses sparse and linear attention to accelerate DiTs.
- **VSA** Zhang et al. (2025d): a trainable sparse attention paradigm that learns adaptive sparse patterns over spatio-temporal tokens.
- **SVG2** Yang et al. (2025a): a dynamic sparse attention method with an adaptive token budget.

Metrics. For video quality, we evaluate on VBench Huang et al. (2023) and report five key dimensions: Subject Consistency (SC), Motion Smoothness (MS), Background Consistency (BC), Imaging Quality (IQ), and Aesthetic Quality (AQ). We additionally report PSNR, SSIM, and LPIPS as complementary frame-level fidelity metrics. Efficiency is measured by DiT end-to-end latency and attention sparsity. Latency denotes the end-to-end runtime of the DiT denoising component, excluding text encoding and VAE stages.

4.3 Experimental Results Analysis

Table 2: Quantitative comparison on Wan2.1-1.3B and Wan2.1-14B benchmarks. We use NVIDIA H100 80G GPUs for all experiments. Wan2.1-1.3B uses 480p inputs, and Wan2.1-14B uses 720p inputs with 81 frames. Latency is measured for the DiT denoising component under the VBench evaluation setting, excluding text encoding and VAE stages. PSNR/SSIM/LPIPS are measured against full-attention outputs with the same random seed.

Model	Method	VBench Quality					Additional Metrics			Efficiency	
		SC↑	MS↑	BC↑	IQ↑	AQ↑	PSNR↑	SSIM↑	LPIPS↓	Latency↓	Sparsity↑
Wan2.1-1.3B	Full	**0.9403**	0.9557	**0.9498**	**0.6510**	**0.6088**	–	–	–	50.9s	0%
	SVG2	0.9201	0.9577	0.9480	0.6319	0.5710	23.12	0.801	0.139	39.7s	72%
	VSA	0.9364	0.9757	0.9421	0.6328	0.5629	23.05	0.819	0.149	34.3s	90%
	SLA	0.9401	0.9782	0.9486	0.6343	0.5918	24.17	0.844	0.130	31.2s	90%
	RoPeSLR (Ours)	**0.9512**	0.9789	0.9485	0.6481	0.6010	25.01	0.843	0.129	**30.4s**	90%
Wan2.1-14B	Full	**0.9698**	0.9850	**0.9681**	**0.6996**	**0.6319**	–	–	–	1075.7s	0%
	SVG2	0.9478	0.9766	0.9501	0.6101	0.5877	23.50	0.832	0.150	789.3s	80%
	VSA	0.9010	0.9788	0.9482	0.6012	0.5998	24.27	0.843	0.130	717.1s	90%
	SLA	0.9481	0.9802	0.9511	0.6398	0.6052	26.02	0.878	0.124	601.0s	90%
	RoPeSLR (Ours)	0.9601	0.9842	0.9528	0.6627	0.6102	26.33	0.880	0.118	**598.3s**	90%

Quality Evaluation. As shown in Table 2, under high-sparsity settings RoPeSLR substantially reduces end-to-end DiT latency relative to dense full attention while remaining competitive on the reported VBench dimensions. The additional PSNR, SSIM, and LPIPS results also indicate strong frame-level agreement with the full-attention outputs, which is a common practice for measuring sparse attention approaches. Overall, these results support a favorable **quality–efficiency trade-off** under the reported evaluation settings.

Latency Characteristics and Sparse Branch Flexibility. For compatibility, we adopt the energy-driven block-sparse branch as the default sparse branch. RoPeSLR is agnostic to this particular choice, and the sparse branch can be replaced by other sparse attention designs. Additional latency analyses are provided in Figure 8, Table 6, and Appendix D.

Supplementary Analyses. We provide additional validations: (i) token-wise gate distributions in Appendix C.5; (ii) the sparse/low-rank profile of LLM attention for comparison in Appendix C.4;

[1]The sparsity ratios of baseline methods follow the default settings of their official implementations. Since some sparse methods do not support exact sparsity control, all reported sparsity values are within a deviation of 1%.

(iii) training hyperparameters in Appendix C.8; (iv) compatibility with step-distilled models for compounded acceleration in Appendix C.7.

4.4 Ablation Study

Experimental Settings. All ablations are conducted on Wan2.1-1.3B with the same settings as the main experiments. Core ablations are consolidated in Table 3. We report ablations under the same training budget and focus on consistent relative trends across configurations.

Low-Rank Compensator Validation. Table 3 compares the sparse-only baseline with the full sparse–low-rank configuration at 90% sparsity. Adding the Fourier low-rank compensator improves all five reported VBench dimensions, indicating that the global low-rank branch helps recover generation quality lost under aggressive sparsification.

Additional Ablation Studies. Appendix ablations include sparsity (Table 4), where 90% is a favorable operating point in our setting; the activation function of the low-rank compensator (Table 5), where a bounded smooth activation performs best; and the feature dimension r (Appendix C.2.2), from which we select $r = 64$ as an empirical operating point.

Table 3: Ablation of the low-rank compensator on Wan2.1-1.3B at 90% sparsity.

Method	SC	MS	BC	IQ	AQ
Sparse only	0.8496	0.9580	0.8909	0.5294	0.4740
Sparse + Low-Rank Compensator (Ours)	**0.9512**	**0.9789**	**0.9485**	**0.6481**	**0.6010**

5 Conclusion and Limitation

Conclusion. We introduce RoPeSLR, an attention framework that targets the $\mathcal{O}(L^2)$ bottleneck in DiTs. Our structural analysis identifies a sparse-plus-low-rank organization in the attention maps of 3D RoPE-equipped DiTs and motivates a Fourier low-rank compensator that preserves relative spatiotemporal offsets. RoPeSLR instantiates this design with an energy-driven block-sparse branch, a compact global low-rank branch, and lightweight gated fusion. Experiments on Wan2.1-1.3B and Wan2.1-14B show that RoPeSLR accelerates inference at 90% sparsity while keeping generation quality close to full attention.

Limitations and future work. The structural analysis provides a design principle rather than a certificate for every deployed component. We instantiate the analysis with hardware-friendly block sparsity, a learnable positive feature map, and gated fusion, validating these choices empirically. Two directions are promising: extending the analysis from attention-matrix entries to attention outputs, and exploring adaptive $\sqrt{L}$ sparse budgets for sub-quadratic inference. On the systems side, integrating stronger sparse kernels should further improve wall-clock speedups.

AI Use Statement

Generative AI tools were used only for language editing and consistency checks, not for the core research contributions, algorithm design, experiments, or results interpretation. The authors reviewed all changes and take full responsibility for the manuscript.

Reproducibility Statement

We provide the complete algorithm, experimental hyperparameters, two-stage post-training procedure, architecture settings ($r = 64$ and 90% sparsity), and evaluation setup in the main paper and appendix. All experiments use the publicly available Wan2.1 models and VBench benchmark.

Ethics Statement

This work improves video-DiT efficiency and does not alter the underlying models' content capabilities. Efficient generation may have dual-use implications; deployments should retain appropriate content filtering and watermarking.

REFERENCES

Pengtao Chen, Xianfang Zeng, Maosen Zhao, Peng Ye, Mingzhu Shen, Wei Cheng, Gang Yu, and Tao Chen. Sparse-vdit: Unleashing the power of sparse attention to accelerate video diffusion transformers, 2025a.

Ruichen Chen, Keith G. Mills, Liyao Jiang, Chao Gao, and Di Niu. Re-ttention: Ultra sparse visual generation via attention statistical reshape, 2025b. URL https://arxiv.org/abs/2505.22918.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andrea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. In *ICLR 2021*, 2021.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL https://arxiv.org/abs/2307.08691.

Qihang Fan, Huaibo Huang, and Ran He. Breaking the low-rank dilemma of linear attention, 2024. URL https://arxiv.org/abs/2411.07635.

Andi Han, Jiaxiang Li, Wei Huang, Mingyi Hong, Akiko Takeda, Pratik Jawanpuria, and Bamdev Mishra. Sltrain: a sparse plus low-rank approach for parameter and memory efficient pretraining, 2024. URL https://arxiv.org/abs/2406.02214.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models, 2023. URL https://arxiv.org/abs/2311.17982.

Boao Kong, Junzhu Liang, Yuxi Liu, Renjia Deng, and Kun Yuan. Cr-net: Scaling parameter-efficient training with cross-layer low-rank structure, 2025. URL https://arxiv.org/abs/2509.18993.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2024. URL https://arxiv.org/abs/2412.03603.

Yuxi Liu, Yipeng Hu, Zekun Zhang, Kunze Jiang, and Kun Yuan. Mixture of distributions matters: Dynamic sparse attention for efficient video diffusion transformers, 2026.

Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. Vdt: General-purpose video diffusion transformers via mask modeling. *arXiv preprint arXiv:2305.13311*, 2023.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Dor Shmilovich, Tony Wu, Aviad Dahan, and Yuval Domb. Liteattention: A temporal sparse attention for diffusion transformers, 2025. URL https://arxiv.org/abs/2511.11062.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Xin Tan, Yuetao Chen, Yimin Jiang, Xing Chen, Kun Yan, Nan Duan, Yibo Zhu, Daxin Jiang, and Hong Xu. Dynamic sparsity in large-scale video dit training. In *Proceedings of the 31st ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*, pp. 101–116, 2026.

Kling Team. Kling-omni technical report, 2025. URL https://arxiv.org/abs/2512.16776.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

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

Jianzong Wu, Liang Hou, Haotian Yang, Tao Xin, Ye Tian, Pengfei Wan, Di ZHANG, and Yunhai Tong. Vmoba: Mixture-of-block attention for video diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 132534–132548, 2026.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, Jianfei Chen, Ion Stoica, Kurt Keutzer, and Song Han. Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. In *International Conference on Machine Learning (ICML)*. PMLR, 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, 2025. URL https://arxiv.org/abs/2502.21079.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers, 2024. URL https://arxiv.org/abs/2410.10629.

Shuo Yang, Haocheng Xi, Yilong Zhao, Muyang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chenfeng Xu, Kelly Peng, Jianfei Chen, Song Han, Kurt Keutzer, and Ion Stoica. Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation, 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations (ICLR)*, 2025b.

Jintao Zhang, Haoxu Wang, Kai Jiang, Shuo Yang, Kaiwen Zheng, Haocheng Xi, Ziteng Wang, Hongzhou Zhu, Min Zhao, Ion Stoica, Joseph E. Gonzalez, Jun Zhu, and Jianfei Chen. Sla: Beyond sparsity in diffusion transformers via fine-tunable sparse-linear attention, 2025a. URL https://arxiv.org/abs/2509.24006.

Jintao Zhang, Jia Wei, Pengle Zhang, Jun Zhu, and Jianfei Chen. Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. In *International Conference on Learning Representations*, volume 2025, pp. 71566–71585, 2025b.

Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. Spargeattention: Accurate and training-free sparse attention accelerating any model inference, 2025c. URL https://arxiv.org/abs/2502.18137.

Peiyuan Zhang, Yongqi Chen, Haofeng Huang, Will Lin, Zhengzhong Liu, Ion Stoica, Eric Xing, and Hao Zhang. Vsa: Faster video diffusion with trainable sparse attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025d.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. URL https://arxiv.org/abs/2412.20404.

Appendix

CONTENTS

1 Introduction **1**

2 Preliminary **2**

3 Method **3**

3.1 Structural Analysis for Design 3

3.2 Fourier Low-Rank Compensator 5

3.3 Sparse-Low-Rank Attention 6

3.4 Gated Fusion 7

3.5 Post-training Strategy 7

4 Experiments **7**

4.1 Experimental Settings 7

4.2 Results and Analysis 8

4.3 Experimental Results Analysis 8

4.4 Ablation Study 9

5 Conclusion and Limitation **9**

A Related work **15**

B Theory **15**

B.1 Proof Overview 15

B.2 Preliminaries and Notation 16

B.3 Core Assumptions 16

B.4 Core Lemmas 17

B.5 Main Proposition 21

C Additional Experiments **24**

C.1 Empirical Validation of Spectral Concentration (Assumption 2) 24

C.1.1 Spectral Evolution and Justification of Exponential Decay 24

C.2 Ablation Study 25

C.2.1 Ablation Study on Sparsity 25

C.2.2 Ablation Study on Low-Rank Dimension 25

C.2.3 Ablation Study on Activation Function 25

C.2.4 Ablation Study on Training Schedule Robustness 27

C.3 Efficiency Analysis 27

C.3.1 Efficiency Analysis of the Low-Rank compensator 27
C.3.2 Efficiency Analysis of Latency for Long-Sequence Video Tokens 28
C.4 Sparse and Low-Rank Characteristics of LLM Attention 28
C.5 Visualization of Learned Gate Distributions 28
C.6 Mechanistic Interpretability via Gram Spectral Analysis 29
C.7 Compatibility with Step-Distilled Models 30
C.8 Algorithm and Hyperparameters 31

D Computational Complexity Analysis **32**

E Additional Video Generation Examples **32**

A Related work

Diffusion Transformers and Efficiency Challenges. Diffusion Transformers (DiTs) Peebles & Xie (2023) have emerged as the dominant architecture for high-fidelity video generation. Earlier decoupled approaches Ma et al. (2024); Lu et al. (2023); Zheng et al. (2024) split video into separate spatial and temporal attention branches, trading joint spatio-temporal modeling capacity for efficiency. Recent unified 3D DiTs Yang et al. (2025b); Kong et al. (2024); Team (2025) process the full token sequence jointly, achieving superior quality but suffering from prohibitive $O(N^2)$ self-attention complexity that scales poorly with increasing video resolution and length.

Sparse Attention for Video DiTs. Sparse attention is the most widely adopted approach to address the quadratic complexity, leveraging the inherent sparsity of video attention patterns where each token only attends to a small subset of relevant tokens. Dynamic sparse attention methods adapt attention patterns to input content: Vmoba Wu et al. (2026) proposes mixture-of-block attention for video diffusion. Sparse-vDiT Chen et al. (2025a) exploits recurring structured sparsity patterns selected through an offline search procedure. Sparse VideoGen Xi et al. (2025) uses online profiling to capture dynamic spatial-temporal sparsity, while Sparse VideoGen2 Yang et al. (2025a) combines semantic-aware permutation with dynamic token-budget control. VSA Zhang et al. (2025d) introduces fully trainable sparse attention. Training-free sparse methods offer plug-and-play acceleration: SpargeAttention Zhang et al. (2025c) and the training-free adaptive sparse attention approach Xia et al. (2025) dynamically adjust sparsity without retraining. Mixture of Distributions Attention (MOD-DiT) Liu et al. (2026) realizes sparse attention via online inference-time prediction of attention patterns. Re-ttention Chen et al. (2025b) achieves ultra-sparsity via attention statistical reshaping, LiteAttention Shmilovich et al. (2025) focuses on temporal sparsity, and DSV Tan et al. (2026) exploits dynamic sparsity for training acceleration.

Sparse-Linear Attention Fusion. While sparse attention delivers significant speedups, it suffers from quality degradation at high sparsity levels. To overcome this, recent works combine sparse attention with linear attention, which has linear complexity but complementary approximation properties. SLA Zhang et al. (2025a) first proposed sparse-linear attention fusion for diffusion transformers. Sana Xie et al. (2024) explores pure linear diffusion transformers for efficient image synthesis.

Low-Rank and Hardware Optimizations. Low-rank approximation is a well-established efficiency technique in LLMs. SLTrain Han et al. (2024) combines sparse and low-rank parameterization for efficient LLM pre-training, while CR-Net Kong et al. (2025) exploits cross-layer low-rank residual structure for parameter-efficient pre-training. However, these methods have not been widely adapted to diffusion transformers, and the "low-rank dilemma" of linear attention Fan et al. (2024) remains a key challenge. In parallel, hardware-optimized attention implementations such as FlashAttention-2 Dao (2023) and SageAttention Zhang et al. (2025b) provide speedups by optimizing memory access and numerical precision.

B Theory

B.1 Proof Overview

We provide an analytical existence result for a sparse-plus-low-rank reconstruction of post-softmax attention in 3D RoPE-equipped video DiTs. The result concerns the attention matrix itself and is independent of the deployed forward-pass implementation.

1. An energy threshold τ splits A into a spike set Ω_τ and background set Ω_τ^c; row stochasticity gives at most $\lfloor 1/\tau \rfloor$ spikes per row.
2. The 3D RoPE logit matrix has an exact Fourier expansion whose individual frequency terms have rank at most two.
3. Under Assumption 2, our measured frequency-index concentration permits a low-rank truncation of the background pre-softmax logits. This result motivates the deployed Fourier bottleneck but is separate from the reconstruction guarantee below.
4. FAVOR+ positive random features Choromanski et al. (2021) give an unbiased exponential-kernel estimator. Because these Gaussian features are unbounded, we use their exact second moment, not Hoeffding concentration.

5. Combining this estimator with exact correction on the spike set gives a sequence-length-independent feature rank for fixed norm and relative-error parameters, and an expected entry-wise squared-error guarantee.

B.2 Preliminaries and Notation

We first formalize all symbols and foundational definitions with zero ambiguity, consistent with standard video diffusion transformer literature.

Notation Conventions

- **Video Tokenization**: A video is tokenized into a sequence of spatiotemporal tokens indexed by $p = (t, h, w)$, where t is the temporal index, h, w are spatial indices. The total sequence length is $L = T \cdot H \cdot W$, with T temporal frames, $H \times W$ spatial resolution per frame.
- **Relative Position**: For tokens p, q, the per-axis relative position is $\Delta_k = p_k - q_k$ for $k \in \{t, h, w\}$, where p_k denotes the k-th coordinate of token p.
- **3D Rotary Position Embedding (3D RoPE)**: The head dimension is split as $d_{\text{head}} = d_t + d_h + d_w$, with d_k the dimension allocated to axis k. For axis k, the m-th frequency pair uses the standard exponential schedule:

$$\theta_m^k = 10000^{-2(m-1)/d_k}, \quad m = 1, 2, \ldots, d_k/2$$

  The 2D rotation matrix for axis k, frequency m, and position p_k is:

$$R_k(p_k, m) = \text{Rot}\theta_m^k p_k$$

  The full RoPE rotation matrix $R(p) \in \mathbb{R}^{d_{\text{head}} \times d_{\text{head}}}$ is block-diagonal, with each block corresponding to $R_k(p_k, m)$ for each axis and frequency. By construction, $R(p)$ is orthogonal: $R(p)^\top R(p) = I$.
- **Frequency Subspace Projection**: For token p, axis k, frequency m, we define the 2D projections of query/key vectors onto the corresponding frequency subspace:

$$Q_p^{(k,m)} = \begin{pmatrix} q_{p,2(m-1)+1}^{(k)} \\ q_{p,2m}^{(k)} \end{pmatrix}, \quad K_q^{(k,m)} = \begin{pmatrix} k_{q,2(m-1)+1}^{(k)} \\ k_{q,2m}^{(k)} \end{pmatrix}$$

  where $q_p^{(k)} \in \mathbb{R}^{d_k}$ is the part of q_p allocated to axis k.
- **Attention Matrices**:
  Pre-softmax QK logit matrix: $s_{p,q} = \langle R(p)q_p, R(q)k_q \rangle$ (we omit the standard $1/\sqrt{d_{\text{head}}}$ scaling for clarity; it can be reintroduced as a linear scaling factor without affecting any bounds).
  Post-softmax attention matrix: $A_{p,q} = \frac{\exp(s_{p,q})}{Z_p}$, where $Z_p = \sum_{r=1}^{L} \exp(s_{p,r})$ is the row partition function.
  Spike/Background Sets: For a given energy threshold $\tau > 0$, we define: High-energy spike set: $\Omega_\tau = \{(p, q) \mid A_{p,q} > \tau\}$; Low-energy background set: $\Omega_\tau^c = \{(p, q) \mid A_{p,q} \le \tau\}$.
  Global Error Tolerance: Fixed constant $\mathcal{E} > 0$, the maximum allowable absolute error per element for the low-rank approximation on the background set.

We provide the core assumptions, lemmas, and main theorem statement with full proofs below.

B.3 Core Assumptions

We state all non-trivial assumptions used in this work, with their structural motivation and empirical support discussed in the accompanying analyses.

Assumption 1 (Bounded Query/Key Projection Weights)**.** *The spectral norms of the trainable query and key projection matrices are uniformly bounded by a constant independent of sequence length L and head dimension d_{head}:*

$$\|W_q\|_2 \le W_{max}, \quad \|W_k\|_2 \le W_{max}$$

where $W_{max} > 0$ is a fixed constant.

Assumption 2 (Spectral Concentration of Background Interaction Coefficients)**.** *For the background set of a **pretrained DiT model** $\Omega_\tau^c = \{(p, q) \mid A_{p,q} \le \tau\}$, the tail interaction coefficients decay*

exponentially with the RoPE pair index for all three axes:

$$\sum_{m=m_0}^{d_k/2} \max_{(p,q)\in\Omega_\tau^c} \left(|a_{p,m}^k| + |b_{p,m}^k|\right) \le C_\omega \cdot \rho_k^{m_0}, \quad \forall m_0 \ge 1$$

where $\rho_k = 10000^{-\alpha/d_k} < 1$, $C_\omega > 0, \alpha > 0$ are constants independent of sequence length L and head dimension d_{head}. Here, $a_{p,m}^k$ and $b_{p,m}^k$ are the Fourier coefficients from the 3D RoPE expansion (Lemma 2).

Remark 1 (Theoretical Grounding and Empirical Validity of Assumption 2)**.** *Assumption 2 is a structural prior for the macroscopic background continuum rather than a universal law. RoPE exhibits a long-term decay property in inter-token dependencies as relative distance increases Su et al. (2021), and self-attention matrices can often be approximated by low-rank matrices Wang et al. (2020). Neither result implies frequency-index decay for our residual background. That property is instead an empirical observation of this work.*

*Its practical plausibility is supported by our decoupled architecture and measurements. Explicitly isolating high-energy interactions into the sparse branch leaves a smoother background set Ω_τ^c. In **Appendix C.1 (Figure 5)**, QK interaction magnitudes decay consistently with frequency index across the evaluated DiTs. Our stability analysis (**Figure 3**) further shows bounded, sub-linear stable-rank growth of the residual background over the tested sequence lengths. These measurements motivate the compact Fourier bottleneck used by RoPeSLR without claiming that the observed decay follows from prior RoPE or low-rank-attention theory.*

B.4 Core Lemmas

Lemma 1 (Deterministic Q/K Boundedness)**.** *By Assumption 1, for any diffusion time step s, the query and key vectors are deterministically bounded across all tokens:*

$$\max_p \left(\|q_p\|_2, \|k_p\|_2\right) \le Q_{max} = W_{max}\sqrt{d_{\text{head}}}$$

where Q_{max} is a constant independent of sequence length L.

Proof. **Step 1: Deterministic Norm Bound for LayerNorm Output.** By definition of LayerNorm, for any token feature vector $x_{s,p}$, we have:

$$\text{LN}(x_{s,p}) = \frac{x_{s,p} - \mu}{\sigma}, \quad \sigma^2 = \frac{1}{d_{\text{head}}} \sum_{i=1}^{d_{\text{head}}} (x_{s,p,i} - \mu)^2$$

The squared ℓ_2 norm of the LayerNorm output is:

$$\|\text{LN}(x_{s,p})\|_2^2 = \sum_{i=1}^{d_{\text{head}}} \left(\frac{x_{s,p,i} - \mu}{\sigma}\right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{d_{\text{head}}} (x_{s,p,i} - \mu)^2$$

By definition of σ^2, $\sum_{i=1}^{d_{\text{head}}} (x_{s,p,i} - \mu)^2 = d_{\text{head}}\sigma^2$. Substituting this in, we have

$$\|\text{LN}(x_{s,p})\|_2^2 = \frac{1}{\sigma^2} \cdot d_{\text{head}}\sigma^2 = d_{\text{head}}$$

Taking the square root of both sides gives the deterministic norm bound:

$$\|\text{LN}(x_{s,p})\|_2 = \sqrt{d_{\text{head}}}$$

Step 2: Norm Bound for Query/Key Vectors. By definition, the query vector is $q_p = W_q \cdot \text{LN}(x_{s,p})$. We use the submultiplicativity property of the spectral norm: for any matrix W and vector x, $\|Wx\|_2 \le \|W\|_2\|x\|_2$. Applying this property:

$$\|q_p\|_2 = \|W_q \cdot \text{LN}(x_{s,p})\|_2 \le \|W_q\|_2 \cdot \|\text{LN}(x_{s,p})\|_2$$

Substituting the bounds $\|W_q\|_2 \le W_{\text{max}}$ and $\|\text{LN}(x_{s,p})\|_2 = \sqrt{d_{\text{head}}}$:

$$\|q_p\|_2 \le W_{\text{max}}\sqrt{d_{\text{head}}}$$

The identical derivation applies to the key vector k_p, since $\|W_k\|_2 \leq W_{\max}$. Since this bound holds for all tokens p, we have:

$$\max_p \left(\|q_p\|_2, \|k_p\|_2\right) \leq W_{\max}\sqrt{d_{\text{head}}}$$

This completes the proof.[2] □

Lemma 2 (3D RoPE Induces Exact Fourier Series)**.** *With standard 3D RoPE, the QK inner product can be written exactly as:*

$$s_{p,q} = \sum_{k\in\{t,h,w\}} \sum_{m=1}^{d_k/2} \left(a^k_{p,m}\cos(\theta^k_m \Delta_k) + b^k_{p,m}\sin(\theta^k_m \Delta_k)\right)$$

where $\Delta_k = p_k - q_k$, and the coefficients are defined as:

$$a^k_{p,m} = Q^{(k,m),1}_p K^{(k,m),1}_q + Q^{(k,m),2}_p K^{(k,m),2}_q, \quad b^k_{p,m} = Q^{(k,m),1}_p K^{(k,m),2}_q - Q^{(k,m),2}_p K^{(k,m),1}_q$$

and satisfy the bounds:

$$|a^k_{p,m}| \leq \|Q^{(k,m)}_p\|_2 \|K^{(k,m)}_q\|_2, \quad |b^k_{p,m}| \leq \|Q^{(k,m)}_p\|_2 \|K^{(k,m)}_q\|_2$$

Proof. **Step 1: Simplify the RoPE QK Inner Product.** By definition, the pre-softmax logit is:

$$s_{p,q} = \langle R(p)q_p, R(q)k_q\rangle$$

Since $R(p)$ is orthogonal, $R(p)^\top R(p) = I$. We use the standard RoPE identity for relative position encoding:

$$\langle R(p)q_p, R(q)k_q\rangle = \left\langle q_p, R(p)^\top R(q)k_q\right\rangle = \langle q_p, R(q-p)k_q\rangle$$

This follows directly from the orthogonality of RoPE rotation matrices.

Step 2: Block-Diagonal Expansion of the Rotation Matrix. The RoPE rotation matrix $R(\Delta)$ (where $\Delta = p - q$) is block-diagonal, with each block corresponding to a 2D rotation matrix for a specific axis and frequency. Thus, the inner product decomposes into a sum over axes and frequencies:

$$\langle q_p, R(\Delta)k_q\rangle = \sum_{k\in\{t,h,w\}} \sum_{m=1}^{d_k/2} \left\langle Q^{(k,m)}_p, R_k(\Delta_k, m)K^{(k,m)}_q\right\rangle$$

where $Q^{(k,m)}_p, K^{(k,m)}_q$ are the 2D frequency subspace projections, and $R_k(\Delta_k, m) = \text{Rot}\theta^k_m\Delta_k$ is the 2D rotation matrix for axis k, frequency m, relative position Δ_k.

Step 3: Explicit 2D Rotation and Fourier Series Form. We expand the 2D inner product explicitly. Let $Q^{(k,m)}_p = (u_1, u_2)^\top$ and $K^{(k,m)}_q = (v_1, v_2)^\top$. The 2D rotation matrix is:

$$R_k(\Delta_k, m) = \begin{pmatrix}\cos(\theta^k_m\Delta_k) & -\sin(\theta^k_m\Delta_k)\\ \sin(\theta^k_m\Delta_k) & \cos(\theta^k_m\Delta_k)\end{pmatrix}$$

Multiplying the rotation matrix by $K^{(k,m)}_q$:

$$R_k(\Delta_k, m)K^{(k,m)}_q = \begin{pmatrix}v_1\cos(\theta^k_m\Delta_k) - v_2\sin(\theta^k_m\Delta_k)\\ v_1\sin(\theta^k_m\Delta_k) + v_2\cos(\theta^k_m\Delta_k)\end{pmatrix}$$

Taking the inner product with $Q^{(k,m)}_p$:

$$\begin{aligned}\left\langle Q^{(k,m)}_p, R_k(\Delta_k, m)K^{(k,m)}_q\right\rangle = {} & u_1\left(v_1\cos(\theta^k_m\Delta_k) - v_2\sin(\theta^k_m\Delta_k)\right)\\ & + u_2\left(v_1\sin(\theta^k_m\Delta_k) + v_2\cos(\theta^k_m\Delta_k)\right)\end{aligned}$$

[2] In practice, LayerNorm applies a learned affine transformation $\gamma \odot \hat{x} + \beta$. We assume these learned scaling and shifting parameters are absorbed into the bounded query/key projection matrices W_q and W_k without loss of generality, maintaining the rigor of the deterministic norm bound.

Grouping terms by $\cos(\theta_m^k \Delta_k)$ and $\sin(\theta_m^k \Delta_k)$:

$$= \underbrace{(u_1 v_1 + u_2 v_2)}_{a_{p,m}^k} \cos(\theta_m^k \Delta_k) + \underbrace{(u_1 v_2 - u_2 v_1)}_{b_{p,m}^k} \sin(\theta_m^k \Delta_k)$$

This gives the exact Fourier series form.

Step 4: Coefficient Bounds. We prove the bounds for $a_{p,m}^k$ and $b_{p,m}^k$ using the Cauchy-Schwarz inequality. For $a_{p,m}^k$:

$$|a_{p,m}^k| = |u_1 v_1 + u_2 v_2| = |\left\langle Q_p^{(k,m)}, K_q^{(k,m)} \right\rangle| \leq \|Q_p^{(k,m)}\|_2 \|K_q^{(k,m)}\|_2$$

For $b_{p,m}^k$, using the same method:

$$|b_{p,m}^k| \leq \|Q_p^{(k,m)}\|_2 \|K_q^{(k,m)}\|_2$$

This completes the proof. □

Lemma 3 (Low-Rank Approximation of Background Pre-Softmax Matrix)**.** *Under Assumption 1 and 2, for any error tolerance $\delta > 0$, there exists a truncated matrix $\tilde{S} \in \mathbb{R}^{L \times L}$ of rank at most:*

$$r = O\left(d_{\text{head}} \log\left(\frac{1}{\delta}\right)\right)$$

such that the approximation error is bounded uniformly over the background set:

$$\max_{(p,q) \in \Omega_\tau^c} |s_{p,q} - \tilde{s}_{p,q}| \leq \delta$$

Proof. We construct the truncated matrix $\tilde{S}$ by keeping only the leading frequency pairs for each axis and prove that the tail error is bounded.

Step 1: Fourier Decomposition and Truncation Strategy. From Lemma 2, the pre-softmax logit is:

$$s_{p,q} = \sum_{k \in \{t,h,w\}} \sum_{m=1}^{d_k/2} T_{k,m}(p,q)$$

where $T_{k,m}(p,q) = a_{p,m}^k \cos(\theta_m^k \Delta_k) + b_{p,m}^k \sin(\theta_m^k \Delta_k)$.

We truncate each axis at frequency index M_k, keeping only the leading frequencies $m \in [1, M_k]$. The truncated matrix is:

$$\tilde{s}_{p,q} = \sum_{k \in \{t,h,w\}} \sum_{m=1}^{M_k} T_{k,m}(p,q)$$

The tail error for each axis k is:

$$\text{Tail}_k(p,q) = \sum_{m=M_k+1}^{d_k/2} T_{k,m}(p,q)$$

Step 2: Bound the Tail Error Using Spectral Concentration. We need to bound the tail error uniformly over the background set Ω_τ^c. Fix an axis k. From Assumption 2, the coefficients decay exponentially:

$$\sum_{m=M_k+1}^{d_k/2} \max_{(p,q) \in \Omega_\tau^c} \left(|a_{p,m}^k| + |b_{p,m}^k|\right) \leq C_\omega \cdot \rho_k^{M_k}$$

From Lemma 2, each frequency term's entry is bounded by:

$$|T_{k,m}(p,q)| \leq |a_{p,m}^k| + |b_{p,m}^k|$$

since $|\cos(\cdot)| \leq 1$ and $|\sin(\cdot)| \leq 1$. Thus, the maximum tail error for axis k is:

$$\max_{(p,q)\in\Omega_\tau^c} |\text{Tail}_k(p,q)| \leq C_\omega \cdot \rho_k^{M_k}$$

Step 3: Choose Truncation Index to Meet Error Budget. We split the error equally across the three axes: we require the tail error for each axis to be at most $\delta/3$. For axis k, we need:

$$C_\omega \cdot \rho_k^{M_k} \leq \frac{\delta}{3}$$

Taking logarithms:

$$M_k \geq \frac{\log(C_\omega) - \log(\delta/3)}{\log(1/\rho_k)} = O\left(\log\left(\frac{1}{\delta}\right)\right)$$

Step 4: Compute the Rank of the Truncated Matrix. From Lemma 4, each frequency matrix $T_{k,m}$ has rank at most 2. The truncated matrix is a sum of $M_t + M_h + M_w$ rank-2 matrices:

$$\tilde{S} = \sum_{k\in\{t,h,w\}} \sum_{m=1}^{M_k} T_{k,m}$$

By the subadditivity of rank:

$$\text{rank}(\tilde{S}) \leq 2(M_t + M_h + M_w) = O\left(\sum_k M_k\right) = O\left(d_{\text{head}} \log\left(\frac{1}{\delta}\right)\right)$$

since the head dimension is split as $d_{\text{head}} = d_t + d_h + d_w$.

Step 5: Verify Uniform Error Bound. The total error is bounded by the sum of tail errors across all axes:

$$\max_{(p,q)\in\Omega_\tau^c} |s_{p,q} - \tilde{s}_{p,q}| \leq \sum_{k\in\{t,h,w\}} \max_{(p,q)\in\Omega_\tau^c} |\text{Tail}_k(p,q)| \leq 3 \cdot \frac{\delta}{3} = \delta$$

This completes the proof. □

Lemma 4 (Each Frequency Term Corresponds to a Rank $\leq$ 2 Matrix)**.** *For any fixed axis $k \in \{t,h,w\}$ and frequency m, the matrix $T_{k,m} \in \mathbb{R}^{L\times L}$ representing the m-th frequency interaction is defined by:*

$$T_{k,m}(p,q) = \left\langle R_k(p_k,m)Q_p^{(k,m)}, R_k(q_k,m)K_q^{(k,m)} \right\rangle$$

This matrix has rank at most 2.

Proof. **Step 1: Define Rotated Query/Key Vectors.** For each token p, define the rotated 2D query vector:

$$\hat{q}_{p,k,m} = R_k(p_k,m)Q_p^{(k,m)} \in \mathbb{R}^2$$

For each token q, define the rotated 2D key vector:

$$\hat{k}_{q,k,m} = R_k(q_k,m)K_q^{(k,m)} \in \mathbb{R}^2$$

By definition, the (p,q)-th entry of $T_{k,m}$ is exactly the Euclidean inner product of these two vectors:

$$T_{k,m}(p,q) = \left\langle \hat{q}_{p,k,m}, \hat{k}_{q,k,m} \right\rangle = \hat{q}_{p,k,m}^\top \hat{k}_{q,k,m}$$

Step 2: Construct Matrix Factorization. We construct two matrices from the rotated vectors: $U_{k,m} \in \mathbb{R}^{L\times 2}$: the p-th row of $U_{k,m}$ is $\hat{q}_{p,k,m}^\top$ (the transpose of the rotated query vector for token p). $V_{k,m} \in \mathbb{R}^{L\times 2}$: the q-th row of $V_{k,m}$ is $\hat{k}_{q,k,m}^\top$ (the transpose of the rotated key vector for token q).

By construction, the transpose of $V_{k,m}$ is a $2 \times L$ matrix, where the q-th column is $\hat{k}_{q,k,m}$.

Step 3: Verify Matrix Product and Rank Bound. We now compute the product $U_{k,m}V_{k,m}^\top$. The (p,q)-th entry of this product is the dot product of the p-th row of $U_{k,m}$ and the q-th column of $V_{k,m}^\top$, which is exactly:

$$\left(U_{k,m}V_{k,m}^\top\right)_{p,q} = \hat{q}_{p,k,m}^\top \hat{k}_{q,k,m} = T_{k,m}(p,q)$$

Thus, we have the exact matrix factorization:

$$T_{k,m} = U_{k,m}V_{k,m}^\top$$

We now apply Sylvester's Rank Inequality: for matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$, $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$. Here, $U_{k,m} \in \mathbb{R}^{L\times 2}$ and $V_{k,m}^\top \in \mathbb{R}^{2\times L}$, so:

$$\text{rank}(T_{k,m}) = \text{rank}(U_{k,m}V_{k,m}^\top) \leq \min(\text{rank}(U_{k,m}), \text{rank}(V_{k,m}^\top)) \leq 2$$

This completes the proof. □

B.5 Main Proposition

Proposition 2 (Energy Threshold Decomposition and Sparsity Bound)**.** *Fix any threshold $\tau \in (0,1)$. Define the spike set and background set:*

$$\Omega_\tau = \{(p,q) : A_{p,q} > \tau\}, \quad \Omega_\tau^c = \{(p,q) : A_{p,q} \leq \tau\}$$

Let A_{sparse} be the restriction of A to the spike set (zero elsewhere), and A_{bg} be the restriction to the background set:

$$A_{sparse}(p,q) = \begin{cases} A_{p,q} & \text{if } (p,q) \in \Omega_\tau \\ 0 & \text{otherwise} \end{cases}, \quad A_{bg}(p,q) = \begin{cases} A_{p,q} & \text{if } (p,q) \in \Omega_\tau^c \\ 0 & \text{otherwise} \end{cases}$$

The following two results hold deterministically for any $\tau \in (0,1)$:

- ***Sparsity Bound**: The sparse branch A_{sparse} contains at most $\lfloor 1/\tau \rfloor$ non-zero entries per row. The total number of non-zero entries is bounded above by $L \cdot \lfloor 1/\tau \rfloor$.*
  ***Asymptotic Sparsity Analysis**: For any sequence length L, the total number of non-zero entries in the sparse branch is bounded by:*

$$NNZ(A_{sparse}) = O(L \cdot \left\lfloor \frac{1}{\tau} \right\rfloor)$$

  .
- ***Background Element-wise Bound**: The background branch A_{bg} is bounded element-wise: $\|A_{bg}\|_\infty \leq \tau$.*

Proof. We prove each statement sequentially, with explicit justification for every step.

Proof of Statement 1 (Sparsity Bound). Fix any row p. By definition, the set of non-zero entries in row p of A_{sparse} is $\{q : A_{p,q} > \tau\}$.

Since each row of the post-softmax attention matrix sums to 1 (normalization constraint), we have:

$$\sum_{q=1}^{L} A_{p,q} = 1$$

For any entry in the sparse set, $A_{p,q} > \tau$ by definition. Therefore, the sum of entries in the sparse set for row p satisfies:

$$\sum_{q\in\Omega_\tau(p)} A_{p,q} \leq \sum_{q=1}^{L} A_{p,q} = 1$$

where $\Omega_\tau(p) = \{q : (p,q) \in \Omega_\tau\}$ denotes the set of tokens q such that (p,q) is in the spike set.

Let $n_p = |\Omega_\tau(p)|$ be the number of non-zero entries in row p of the sparse branch. Since each entry in the sparse set is strictly greater than τ, we have:

$$\sum_{q \in \Omega_\tau(p)} A_{p,q} > n_p \cdot \tau$$

Combining the two inequalities:

$$n_p \cdot \tau < \sum_{q \in \Omega_\tau(p)} A_{p,q} \leq 1$$

Thus:

$$n_p < \frac{1}{\tau}$$

Since n_p is an integer, we have $n_p \leq \lfloor 1/\tau \rfloor$. This bound holds for every row, so the total number of non-zero entries across all rows is:

$$\text{NNZ}(A_{\text{sparse}}) = \sum_{p=1}^{L} n_p \leq L \cdot \left\lfloor \frac{1}{\tau} \right\rfloor$$

Asymptotic Analysis. For the asymptotic analysis, when $\tau = O(L^{-1/2})$, we have:

$$\left\lfloor \frac{1}{\tau} \right\rfloor = O(L^{1/2})$$

Thus:

$$\text{NNZ}(A_{\text{sparse}}) = O(L \cdot L^{1/2}) = O(L^{3/2})$$

Proof of Statement 2 (Background Element-wise Bound). By definition, the background set is $\Omega_\tau^c = \{(p,q) : A_{p,q} \leq \tau\}$. Thus, every entry in A_{bg} satisfies $A_{\text{bg}}(p,q) \leq \tau$ by construction. Taking the maximum over all entries:

$$\|A_{\text{bg}}\|_\infty = \max_{p,q} |A_{\text{bg}}(p,q)| \leq \tau$$

This completes the proof. □

Lemma 5 (Positive Random Features for the Exponential Kernel Choromanski et al. (2021))**.** *Let $x, y \in \mathbb{R}^{d_{\text{head}}}$ satisfy $\|x\|_2, \|y\|_2 \leq B$. For $\omega_1, \ldots, \omega_R \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$, define*

$$\varphi_\omega(x) = \exp\left(-\frac{\|x\|_2^2}{2}\right) \exp(\omega^\top x), \qquad \Phi_R(x) = \frac{1}{\sqrt{R}} \big(\varphi_{\omega_1}(x), \ldots, \varphi_{\omega_R}(x)\big)^\top.$$

Then $\hat{K}_R(x,y) = \Phi_R(x)^\top \Phi_R(y)$ is a positive, unbiased estimator of $K(x,y) = \exp(x^\top y)$ and satisfies

$$\frac{\mathbb{E}[(\hat{K}_R(x,y) - K(x,y))^2]}{K(x,y)^2} = \frac{\exp(\|x+y\|_2^2) - 1}{R} \leq \frac{\exp(4B^2) - 1}{R}.$$

For any collections $\{x_i\}_{i=1}^L$ and $\{y_j\}_{j=1}^L$, the matrix $\hat{K}$ with entries $\hat{K}_{ij} = \hat{K}_R(x_i, y_j)$ has rank at most R.

Proof. The Gaussian moment-generating identity $\mathbb{E}_\omega[\exp(\omega^\top z)] = \exp(\|z\|_2^2/2)$ gives

$$\mathbb{E}_\omega[\varphi_\omega(x)\varphi_\omega(y)] = \exp(x^\top y) = K(x,y).$$

For $X_\omega = \varphi_\omega(x)\varphi_\omega(y)$, the same identity yields

$$\frac{\mathbb{E}[X_\omega^2]}{K(x,y)^2} = \exp(\|x+y\|_2^2).$$

Averaging R independent copies therefore gives the stated relative mean-squared error, and $\|x+y\|_2 \leq 2B$ gives the inequality. These log-normal features are unbounded; the argument uses their finite

second moment and does not invoke Hoeffding's inequality. Finally, $\hat{K} = \Phi_X \Phi_Y^\top$ for two $L \times R$ feature matrices, so $\text{rank}(\hat{K}) \le R$. □

Lemma 6 (Exact Diagonal Normalization and Rank Preservation). *Let $Z_p = \sum_{r=1}^{L} \exp(s_{p,r})$ be the true row partition function, and define $D = diag(1/Z_1, \dots, 1/Z_L)$. For any approximated exponentiated kernel matrix $\hat{E} \in \mathbb{R}^{L \times L}$, the normalized matrix $\hat{A} = D\hat{E}$ satisfies:*

$$rank(\hat{A}) = rank(\hat{E})$$

Proof. **Step 1: Prove D is Invertible.** By definition, the partition function $Z_p = \sum_{r=1}^{L} \exp(s_{p,r})$. Since the exponential function is positive ($\exp(x) > 0$ for all $x \in \mathbb{R}$), every term in the sum is positive, so $Z_p > 0$ for all rows p.

The diagonal matrix D has entries $1/Z_p$ on its diagonal, all of which are positive. A diagonal matrix is invertible if and only if all its diagonal entries are non-zero. Thus, D is full-rank and invertible, with inverse $D^{-1} = \text{diag}(Z_1, Z_2, \dots, Z_L)$.

Step 2: Prove Rank Preservation. For any matrix $\hat{E} \in \mathbb{R}^{L \times L}$ and invertible matrix $D \in \mathbb{R}^{L \times L}$, left-multiplying by D preserves the rank:

$$\text{rank}(D\hat{E}) = \text{rank}(\hat{E})$$

Since $\hat{A} = D\hat{E}$, we have:

$$\text{rank}(\hat{A}) = \text{rank}(\hat{E})$$

This completes the proof. □

Corollary 1 (Constructive Sparse-Low-Rank Reconstruction of Attention). *Under Assumption 1, let $C_Q = \exp(4Q_{\max}^2) - 1$. For any $\tau \in (0, 1)$ and $\mathcal{E} > 0$, choose*

$$R = \max\left\{1, \left\lceil C_Q \left(\frac{\tau}{\mathcal{E}}\right)^2 \right\rceil\right\}.$$

There exists a randomized reconstruction $\hat{A}_{\text{final}} = \tilde{A}_{\text{sparse}} + \hat{A}_{\text{lowrank}}$ such that rank$(\hat{A}_{\text{lowrank}}) \le R$, the residual sparse branch has at most $\lfloor 1/\tau \rfloor$ nonzeros per row, and

$$\max_{p,q} \mathbb{E}\left[|\hat{A}_{\text{final}}(p,q) - A(p,q)|^2\right] \le \mathcal{E}^2.$$

The expectation is over the shared Gaussian random features.

Proof. Set $x_p = R(p)q_p$ and $y_q = R(q)k_q$. Orthogonality of RoPE and Lemma 1 give $\|x_p\|_2, \|y_q\|_2 \le Q_{\max}$. Apply Lemma 5 to obtain $\hat{E}_{p,q} = \hat{K}_R(x_p, y_q)$, and let $D = \text{diag}(1/Z_1, \dots, 1/Z_L)$ and $\hat{A}_{\text{lowrank}} = D\hat{E}$. Lemma 6 gives $\text{rank}(\hat{A}_{\text{lowrank}}) \le R$, while unbiasedness and the second-moment bound give

$$\mathbb{E}\left[|\hat{A}_{\text{lowrank}}(p,q) - A(p,q)|^2\right] \le \frac{C_Q}{R} A(p,q)^2.$$

Define

$$\tilde{A}_{\text{sparse}}(p,q) = \begin{cases} A(p,q) - \hat{A}_{\text{lowrank}}(p,q), & (p,q) \in \Omega_\tau, \\ 0, & (p,q) \in \Omega_\tau^c. \end{cases}$$

The error is zero on Ω_τ. On Ω_τ^c, $A(p,q) \le \tau$, so the chosen R gives expected squared error at most $C_Q\tau^2/R \le \mathcal{E}^2$. Proposition 2 gives the claimed sparsity. □

The following proposition is the formal version of Proposition 1.

Proposition 3. *Under Assumption 1, let $C_Q = \exp(4Q_{\max}^2) - 1$, choose $\tau = cL^{-1/2}$ and $\mathcal{E} = \tau/2$, and set $R = \max\{1, \lceil 4C_Q \rceil\}$. Then the randomized reconstruction in Corollary 1 satisfies*

1. $\mathrm{NNZ}(\tilde{A}_{\mathrm{sparse}}) = \mathcal{O}(L^{3/2})$;
2. *rank*$(\hat{A}_{\mathrm{lowrank}}) \le R$, *where* R *is independent of* L *for fixed* $Q_{\max}$;
3. $\max_{p,q} \left(\mathbb{E}|\hat{A}_{\mathrm{final}}(p,q) - A(p,q)|^2\right)^{1/2} \le \frac{c}{2L^{1/2}} = \mathcal{O}(L^{-1/2})$.

Proof. The deterministic sparsity bound gives

$$\mathrm{NNZ}(\tilde{A}_{\mathrm{sparse}}) \le L \left\lfloor \frac{1}{\tau} \right\rfloor = \mathcal{O}(L^{3/2}).$$

Because $\tau/\mathcal{E} = 2$, Corollary 1 requires at most $R = \max\{1, \lceil 4C_Q \rceil\}$ features and gives the stated expected root-mean-squared entry-wise error. □

Remark 2 (Superiority of Sparse-Low-Rank Structure in DiTs and LLMs)**.** *Our framework analyzes why DiTs may exhibit a more pronounced sparse-low-rank structure than LLMs, grounded in two mathematical inequalities:*

1. ***Cubic Acceleration of Spectral Decay (Rank Inequality)****: As derived in Lemma 3, the truncation index M_k required to bound the error depends on the exponential decay constant ρ. For LLMs using 1D RoPE with head dimension d_{head}, the decay rate is $\rho_{LLM} = 10000^{-\alpha/d_{\mathrm{head}}}$. For DiTs using 3D RoPE, the dimension is partitioned (d_t, d_h, d_w). Assuming a uniform partition $d_k = d_{\mathrm{head}}/3$, the per-axis decay rate is $\rho_{DiT} = 10000^{-\alpha/(d_{\mathrm{head}}/3)} = \rho_{LLM}^3$.*
   Recall from the proof of Lemma 3 that the required rank R to achieve an error δ scales inversely with $-\log(\rho)$. Therefore, we have the theoretical relationship:
   $$R_{DiT} \propto \frac{1}{-\log(\rho_{DiT})} = \frac{1}{-3\log(\rho_{LLM})} = \frac{1}{3} R_{LLM} \tag{17}$$
   This suggests that, under a uniform dimension split and identical error tolerance, the stable rank of the background matrix in DiTs is asymptotically one-third that of LLMs, amplifying the efficiency of the low-rank branch.
2. ***Empirical Tightness of the Deterministic Sparsity Bound****: While the theoretical $\mathcal{O}(L^{3/2})$ sparsity upper bound (derived via NNZ $\le L \cdot \lfloor 1/\tau \rfloor$) is a deterministic property holding for any row-stochastic attention matrix, its practical utility heavily depends on the data domain. In LLMs, tasks like in-context learning scatter semantic spikes across the causal history, causing this universal bound to be vacuous in practice. Conversely, the inherent spatiotemporal locality of video naturally clusters high-energy semantic interactions. This physical structure ensures that the theoretical deterministic bound aligns tightly with the empirical attention manifold in DiTs, rendering the $\mathcal{O}(L^{3/2})$ extraction practically effective.*
3. ***Empirical Evidence****: We provide experimental evidence to validate this remark, see figure 2 and 9 for more details.*

C Additional Experiments

C.1 Empirical Validation of Spectral Concentration (Assumption 2)

C.1.1 Spectral Evolution and Justification of Exponential Decay

To assess the exponential decay $\rho_k^{m_0}$ ($\rho_k < 1$) posited in Assumption 2, we measure the contribution of each frequency index m for axis $k \in \{t, h, w\}$. We define the **Robust Interaction Magnitude** $\mathcal{M}(k, m)$ using the 99th percentile:

$$\mathcal{M}(k,m) = \mathrm{Percentile}_{99\%}\left(\left|\langle Q_p^{(k,m)}, R_k(\Delta) K_q^{(k,m)} \rangle\right|\right) \tag{18}$$

A strict maximum over all L^2 pairs is dominated by isolated sparse spikes. To reduce sensitivity to such outliers, we report the 99th percentile as a robust empirical summary of the background interaction magnitude; it is not used as an estimator of the mathematical supremum in Assumption 2. We simultaneously compute the **Cumulative Tail Energy**: $\mathrm{Tail}(m_0) = \sum_{m=m_0}^{d_k/2} \mathcal{M}(k, m)$.

Understanding the Visualizations. We evaluate these metrics on Wan2.1-T2V-1.3B, Wan2.1-T2V-14B, and HunyuanVideo-13B across multiple denoising stages. In Figure 5, the horizontal axes denote the RoPE dimension index m (larger m corresponds to a lower rotational frequency). The vertical axes are on a **logarithmic scale** ($\log_{10}$). An exponential decay $f(m) = C \cdot \rho^m$ translates to a linear downward trend on a semi-log plot.

Key Observations & Generalization. Figure 5 shows that across **all three models**, the interaction energy along the t, h, and w dimensions exhibits a consistent decline on the Log-Y axes as m increases. This observation supports the spectral-concentration prior used in Assumption 2 across the evaluated models, while not establishing it as a universal property of all 3D RoPE-equipped DiTs. Furthermore, as denoising progresses from Gaussian noise to semantic generation, this decay steepens (i.e., ρ_k decreases), providing additional empirical support for the assumption.

C.2 Ablation Study

C.2.1 Ablation Study on Sparsity

We perform an ablation study on sparsity. Since the energy-thresholded block-sparse branch cannot always precisely control the exact sparsity ratio, all sparsity values in the experiments are approximate results with an error within 1%. Results in Table 4 show that increasing sparsity from 80% to 90% induces nearly no quality loss, while a further rise from 90% to 95% causes obvious performance degradation. The sparsity of 90% is therefore a favorable sweet spot.

Table 4: Ablation study on attention sparsity.

Sparsity	SC	MS	BC	IQ	AQ
80%	0.9607	0.9899	0.9502	0.6475	0.6075
90%	0.9512	0.9789	0.9485	0.6481	0.6010
95%	0.9387	0.9763	0.9127	0.6193	0.5870

C.2.2 Ablation Study on Low-Rank Dimension

Figure 6 illustrates the Stage-I alignment dynamics across varying low-rank dimensions on Wan2.1-T2V-1.3B. While all configurations converge reliably, aggressively compressed bottlenecks (e.g., $r = 8$) exhibit higher loss floors and optimization variance, indicating constrained representational capacity. Within the tested range, increasing the bottleneck dimension generally improves approximation fidelity and convergence stability. Among the tested values, $r = 64$ **achieves the lowest MSE loss and the most stable optimization behavior**. Consequently, we use $r = 64$ as our default empirical operating point; it should not be interpreted as a strict theoretical rank bound.

C.2.3 Ablation Study on Activation Function

We further conduct an ablation study on the activation function used in the low-rank compensator. In this experiment, we keep the sparse branch fixed as the energy-driven block-sparse branch and only replace the non-linear activation in the low-rank compensator, so that the effect of different activation choices can be isolated.

As shown in Table 5, Sigmoid achieves the best overall performance among the tested activation functions. It obtains the highest scores on SC, BC, IQ, and AQ, while maintaining a competitive MS score. This indicates that a bounded and smooth activation is more suitable for the low-rank compensator, helping it provide stable complementary information to the sparse attention branch.

Table 5: Comparison of activation functions in the low-rank compensator.

Activation Function	**SC↑**	**MS↑**	**BC↑**	**IQ↑**	**AQ↑**
ReLU	0.9353	0.9871	0.9346	0.6458	0.5689
GELU	0.9371	**0.9884**	0.9395	0.6291	0.5906
Tanh	0.9433	0.9854	0.9414	0.6465	0.5421
Sigmoid	**0.9512**	0.9789	**0.9485**	**0.6481**	**0.6010**

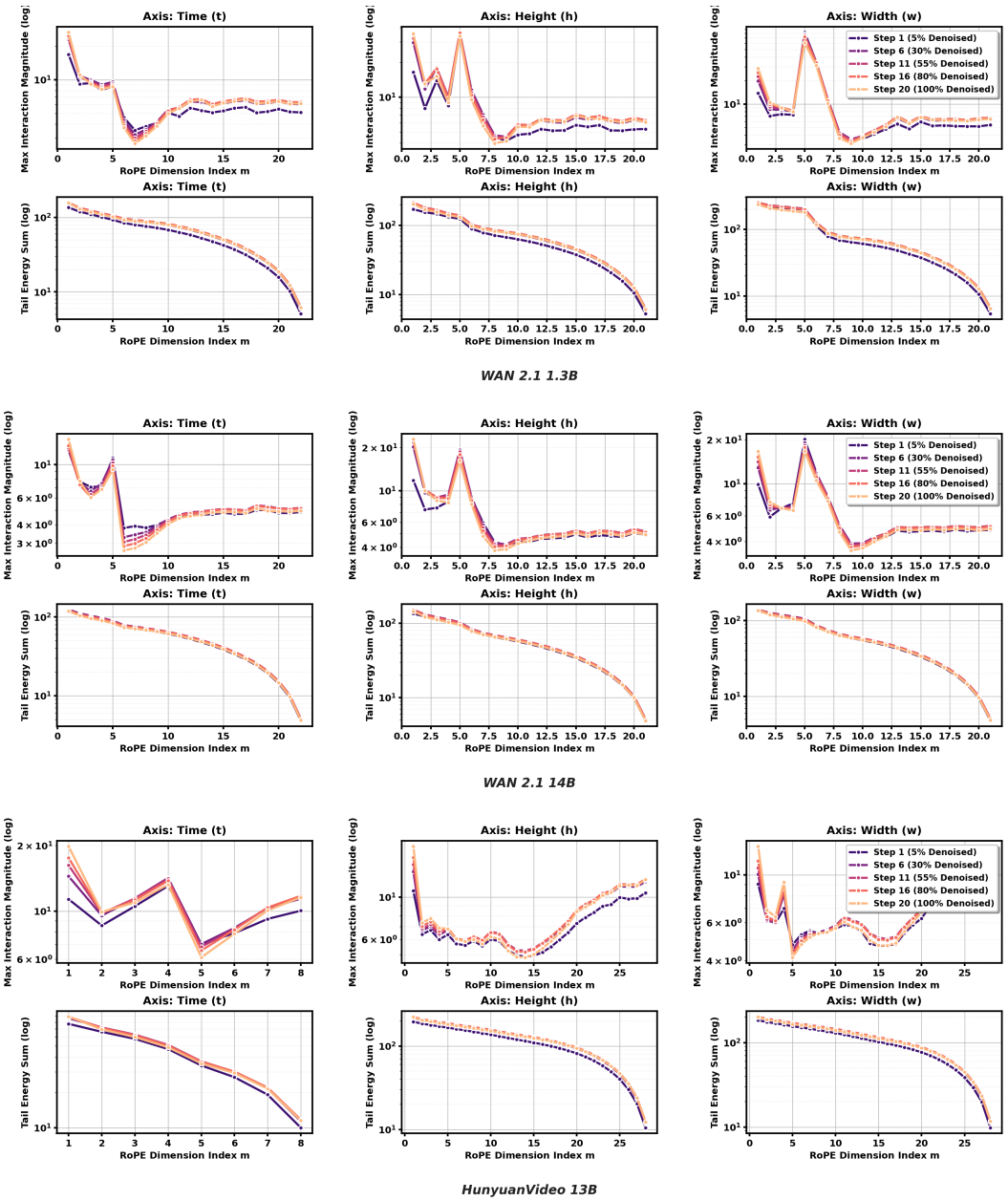

Figure 5: **Evolution of QK Spectral Concentration with 3D RoPE.** We evaluate three pre-trained DiTs (Wan2.1-1.3B, Wan2.1-14B, and HunyuanVideo-13B) using 3 representative text prompts at 480×832 resolution and 49 frames. The X-axis is the RoPE pair index $m \in [1, d_k/2]$. We report the 99th-percentile interaction magnitude and cumulative tail energy on a **Log-Y scale**. The consistent downward trends support Assumption 2 for the evaluated models, not as a universal property.

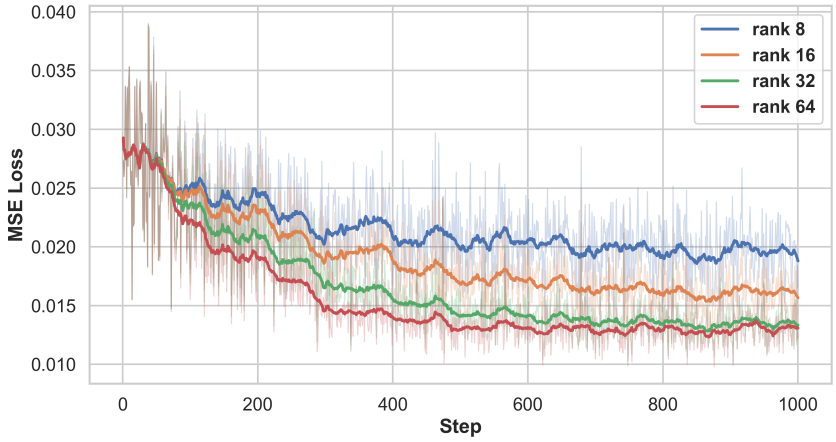

Figure 6: Training MSE loss under different low-rank dimensions.

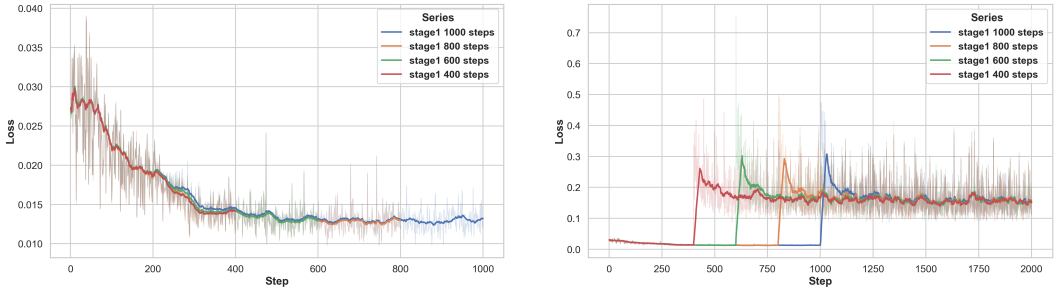

Figure 7: Left: Stage-I training loss under different stage-I lengths. Right: Overall training loss under different stage-I lengths.

C.2.4 Ablation Study on Training Schedule Robustness

To validate the stability of our two-stage post-training pipeline, we analyze the loss dynamics under varying Stage-I alignment durations (ranging from 400 to 1000 steps). As illustrated in Figure 7, the Stage-I MSE loss reliably converges to a shared minimum regardless of the scheduled length, ensuring sufficient structural alignment of the newly introduced low-rank compensator. Furthermore, upon transitioning to Stage-II full fine-tuning, the optimization trajectories swiftly stabilize and converge to comparable optimal levels across all settings. These observations demonstrate that our decoupled training strategy is robust and insensitive to the specific transition threshold.

C.3 Efficiency Analysis

C.3.1 Efficiency Analysis of the Low-Rank Compensator

We analyze the computational overhead of different approximation methods for the lowrank compensator (bottom-90%), comparing our low-rank auxiliary compensator against two baseline implementations: full softmax attention and linear attention.

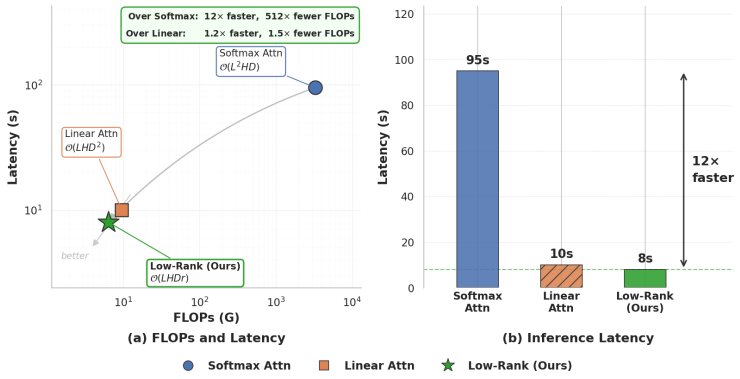

Figure 8: FLOPs and latency comparison of attention approximation methods for the non-top-k branch on Wan2.1-1.3B. (a) FLOPs-latency trade-off on a log-log scale; (b) absolute inference latency per forward pass.

As shown in Figure 8, RoPeSLR delivers the best overall efficiency among the three methods. Compared to full softmax attention on the same background token set, it reduces branch FLOPs by $256\times$ and branch latency by $12\times$. It also outperforms linear attention, delivering $1.25\times$ faster inference with lower computational overhead.

C.3.2 Efficiency Analysis of Latency for Long-Sequence Video Tokens

We further evaluate the latency of RoPeSLR on long-sequence video generation using HunyuanVideo-13B. The inference is conducted with 129 frames at 720×1280 resolution, corresponding to $33 \times 45 \times 80 = 118{,}800$ video tokens after VAE compression and patchification. We compare the end-to-end inference latency of the full-attention baseline and RoPeSLR under the same inference setting, as shown in Table 6.

Table 6: End-to-end latency comparison on HunyuanVideo-13B with 118,800 long-sequence video tokens.

Method	**Latency (s)↓**	**Speedup↑**
Full Attention	3354s	–
RoPeSLR (Ours)	**1482s**	**2.26×**

C.4 Sparse and Low-Rank Characteristics of LLM Attention

We analyze the causal self-attention weight matrices from Qwen2-7B-Instruct (28 GQA layers, 28 query heads per layer). For each row of the attention matrix, we retain the entries with the top 90% cumulative energy and exclude them, then analyze the remaining low-energy residual part. Rows with fewer than 128 visible keys are excluded, resulting in 784 valid (layer, head) combinations.

As illustrated in Figure 9, the left energy concentration curve presents a steeper rising trend at the initial segment compared with video diffusion Transformers, which is mainly attributed to the inherent attention sink effect in LLMs. The right subfigure presents the stable rank distribution of low-energy residuals, with a median stable rank around 40. Compared with video DiT, the substantially higher stable rank of LLM residuals indicates that **LLMs do not exhibit the prominent sparse and low-rank structural patterns observed in video diffusion models** (see Figure 2).

C.5 Visualization of Learned Gate Distributions

We visualize the gate activation intensity across denoising timesteps to examine whether the gating mechanism captures content-dependent spatial structure. Recall that the scalar gate determines how much global Fourier low-rank context is injected into the sparse representation at each token. As shown in Figure 10, we present the gate activation maps for five representative layers at ten uniformly sampled denoising timesteps, alongside a reference frame from the generated video. Two consistent patterns can be observed.

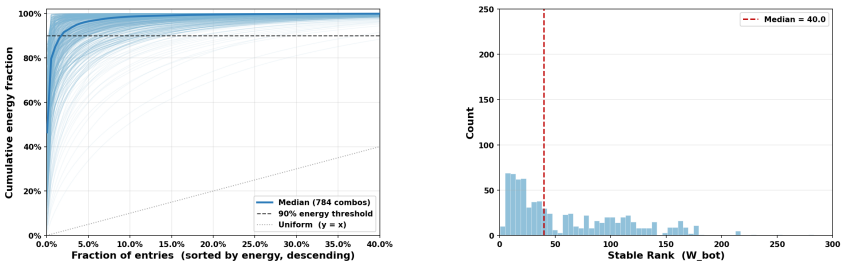

Figure 9: Analysis of causal self-attention matrices from Qwen2-7B-Instruct under a 90% energy threshold over 784 valid combinations.

Spatial structure in gate activation aligns with scene content. The gate activation maps are not spatially uniform but exhibit structural contours that correspond to the scene layout visible in the reference frame: the airplane body, the sky–ground boundary, and the grass–runway edge are all visibly delineated in the gate intensity maps across all five layers. This indicates that the gate responds to the semantic content of the video rather than assigning uniform weight across spatial positions.

Spatial structure becomes more pronounced in later denoising steps. In early denoising steps ($t \leq 11$), gate activation is relatively uniform across spatial positions. As denoising progresses ($t \geq 22$), the spatial contours become increasingly distinct, with higher gate activation concentrated around semantically meaningful regions such as the airplane body and scene boundaries. This evolution suggests that the gate progressively adapts to the emerging content structure as the video is refined through the denoising process.

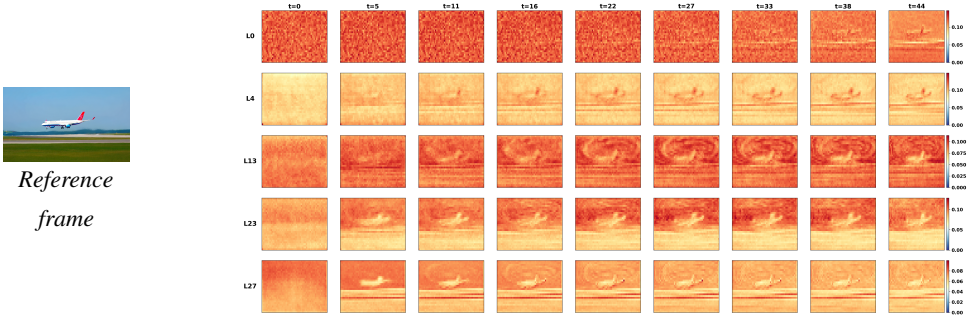

Figure 10: Gate activation intensity maps of our RoPeSLR on Wan2.1-T2V-1.3B (480P), shown for five representative layers (L0, L4, L13, L23, L27) alongside a reference frame from the generated video (prompt: *"an airplane taking off"*). Each row corresponds to one transformer layer; each column is a uniformly sampled denoising timestep (t=0 to t=49, 10 steps). Each cell shows the mean gate value over all attention heads, averaged over T=21 latent frames and projected onto the 30×52 spatial latent grid. Each row uses an independent color scale.

C.6 Mechanistic Interpretability via Gram Spectral Analysis

To mechanistically validate the algebraic insights from the low-rank decomposition, we perform Singular Value Decomposition (SVD) on the pre-fusion sequence-level output of the rank-64 Fourier low-rank branch, $O_{\text{lr}} \in \mathbb{R}^{L \times d_{\text{head}}}$. By factoring $O_{\text{lr}} = U\Sigma V^\top$, we extract the singular values σ_i, which quantify the energy captured by each mode, and the left singular vectors $\mathbf{u}_i \in \mathbb{R}^L$, which represent the spatial principal components. Reshaping $\mathbf{u}_i$ back to the native spatiotemporal grid allows us to directly visualize the learned spatial modes.

Eigenspectrum Analysis: Spectral Collapse and Structural Alignment. The normalized eigenspectrum (Figure 11, Top) plots the energy distribution across the principal components. Without the RoPE-derived positional prior, the low-rank branch exhibits a substantially flatter spectrum, whereas retaining the Fourier positional prior preserves more spatial variance. The resulting spectrum follows the overall trend of the Full Attention oracle and decays rapidly. The sharp energy drop near the practical feature dimension $r = 64$ supports our empirical rank choice, but should not be interpreted as a strict theoretical rank boundary.

Spatial Mode Analysis: Pure Decoupling and Low-pass Filtering. Visualizing the reshaped dominant spatial modes (Figure 11, Bottom) reveals the underlying manifold decoupling mechanism. The principal components of exact Full Attention are visibly entangled with high-frequency, localized semantic spikes, such as sharp object outlines. In stark contrast, the Fourier low-rank branch filters out these local semantics and synthesizes highly pure, low-frequency geometric standing waves. This provides evidence that the low-rank branch acts as a spatial low-pass filter: it decodes the smooth background continuum, while leaving high-frequency semantic routing to the complementary sparse branch.

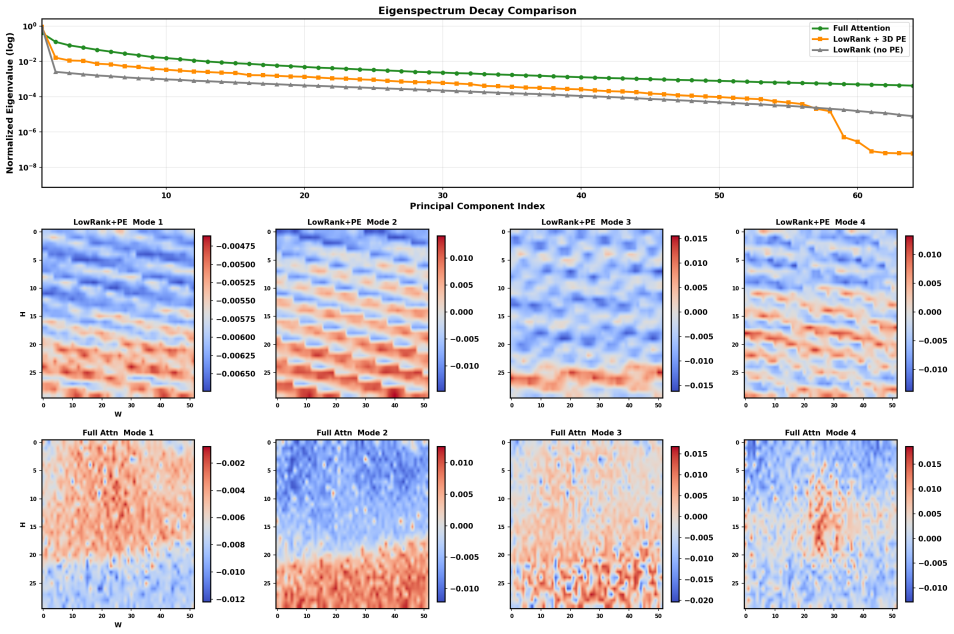

Figure 11: **Mechanistic Interpretability via Gram Spectral Analysis.** *Methodology:* Evaluated on Wan2.1-T2V-1.3B (Layer 15, Denoising Step 10) at 480×832 resolution. SVD is applied to the pre-fusion output O_{lr} of the Fourier low-rank branch over a latent grid of $18 \times 30 \times 52$ ($L = 28{,}080$). *Analysis:* **(Top)** The eigenspectrum with the RoPE-derived positional prior retains more structured spatial variance and follows the overall decay trend of Full Attention, unlike the flatter no-positional-prior baseline. The energy drop near the practical feature dimension $r = 64$ is consistent with our empirical rank choice. **(Bottom)** Heatmaps of the top-4 spatial principal components ($\mathbf{u}_1$ to $\mathbf{u}_4$) at center frame $t = 9$. Unlike the noisy, high-frequency semantics entangled in Full Attention, the Fourier low-rank branch generates clean geometric standing waves, validating its role as a spatial low-pass filter.

C.7 Compatibility with Step-Distilled Models

Our method does not alter the denoising schedule or the external model architecture, so it is architecturally compatible with step-distilled models. In principle, this allows the attention cost reduction to be combined with step distillation. Step distillation reduces the number of diffusion denoising steps, while our method lowers the attention computation cost per step. This compatibility should not be interpreted as a guarantee that the approximation behavior remains identical after distillation.

Why Sparse Routing Remains Effective. RoPeSLR leverages the inherent spatiotemporal sparsity of natural video signals through energy-driven block-sparse routing. Due to the strong redundancy across consecutive video frames, attention energy naturally concentrates in local spatiotemporal neighborhoods. This structural property is determined primarily by input data characteristics, rather than the number of denoising steps.

Step-distilled models use larger denoising jumps per inference step, which may alter the distribution of attention energy. It is therefore plausible that sparse selection could benefit from more concentrated informative regions, but this behavior should be verified empirically for each distilled model.

Transferability of the Low-Rank Branch. The low-rank branch of RoPeSLR is formulated as a global Fourier state readout:

$$O^{(h)}_{\mathrm{lr},p} = \frac{(\phi^h_p)^\top C^h}{(\phi^h_p)^\top u^h + \epsilon}, \qquad C^h = \sum_{q=1}^{L} \psi^h_q (v^h_q)^\top, \qquad u^h = \sum_{q=1}^{L} \psi^h_q.$$

It depends on the per-head feature dimension, the number of attention heads, the retained RoPE frequency set, and the learned positive feature map Φ^h. Step distillation generally preserves the original model architecture, including hidden dimension, number of attention heads, per-head dimension, RoPE frequency schedule, and latent spatiotemporal grid. These shared architectural ingredients make direct application possible, but they do not by themselves guarantee unchanged feature statistics or approximation quality after distillation.

C.8 Algorithm and Hyperparameters

Algorithm 1 RoPeSLR Attention Mechanism

Input: Input sequence $X \in \mathbb{R}^{L \times d_{\mathrm{model}}}$, 3D coordinates (t, h, w), energy threshold τ, low-rank dimension r, retained RoPE frequency set $\mathcal{I}$, and learnable parameters Θ.
Output: Refined representation $O \in \mathbb{R}^{L \times d_{\mathrm{model}}}$.

// Step 1: Dual-branch routing (per head h)

for each attention head $h \in \{1, \ldots, H\}$ **do**

Sparse branch (energy-driven):

Determine the energy-thresholded block-sparse set $\mathcal{S}^h_p$ for each query p.

Compute sparse attention exactly on the retained set:

$$O^{(h)}_{\mathrm{sp},p} = \sum_{q \in \mathcal{S}^h_p} \frac{\exp(s^h_{p,q})}{\sum_{r \in \mathcal{S}^h_p} \exp(s^h_{p,r})} v^h_q.$$

Fourier low-rank branch:

Construct rotated low-frequency Fourier features:

$$z^h_p = R^h_{\mathcal{I}}(p)\, q^h_{p,\mathcal{I}}, \qquad y^h_q = R^h_{\mathcal{I}}(q)\, k^h_{q,\mathcal{I}}.$$

Apply the positive feature map:

$$\phi^h_p = \Phi^h(z^h_p), \qquad \psi^h_q = \Phi^h(y^h_q).$$

Aggregate the global Fourier state:

$$u^h = \sum_{q=1}^{L} \psi^h_q, \qquad C^h = \sum_{q=1}^{L} \psi^h_q (v^h_q)^\top.$$

Read out the low-rank output:

$$O^{(h)}_{\mathrm{lr},p} = \frac{(\phi^h_p)^\top C^h}{(\phi^h_p)^\top u^h + \epsilon}.$$

end for

// Step 2: Gated fusion

Compute the token-wise scalar gate:

$$g = \sigma(X W_g + b_g).$$

Fuse the sparse and low-rank branches and stabilize with RMSNorm:

$$O^{(h)} = \mathrm{RMSNorm}\left(O^{(h)}_{\mathrm{sp}} + g \odot O^{(h)}_{\mathrm{lr}}\right).$$

Concatenate the per-head outputs to obtain O.

return O

Table 7 lists the main hyperparameters used for fine-tuning.

Table 7: Training and Model Hyperparameters for RoPeSLR on Wan2.1-T2V-1.3B

Hyperparameter	**Value**
Optimizer	AdamW
β_1, β_2	0.9, 0.99
Weight decay	10^{-5}
Max gradient norm	1.0
Precision	BF16
Low-rank dimension r	64
Two-Stage Training Schedule	
Stage I loss	MSE
Stage II loss	Flow Matching Loss
Stage I LR schedule	Constant
Stage II LR schedule	Cosine
Learning Rates (Stage I / Stage II)	
Backbone	— / 10^{-5}
Fourier low-rank feature map Φ	3×10^{-4} / 10^{-4}
Gate projection	3×10^{-4} / 10^{-4}
Gate bias	3×10^{-4} / 10^{-4}
PE scale α_{pe}	3×10^{-4} / 10^{-4}

D Computational Complexity Analysis

To evaluate the efficiency of RoPeSLR, we provide a detailed theoretical analysis of the floating-point operations (FLOPs) required during the forward pass. Under identical sparsity, RoPeSLR is structurally more computationally efficient than existing sparse-plus-linear attention frameworks when the low-rank dimension r is small relative to the head dimension d_{head}. Compared to pure sparse attention, the extra computational overhead introduced by the Fourier low-rank branch is asymptotically negligible for long sequences. Detailed derivations are available in the extended version.

E Additional Video Generation Examples

We provide additional qualitative examples comparing full attention baseline with RoPeSLR at 90% sparsity. Each figure shows video frames from different prompts, demonstrating that RoPeSLR maintains visual quality across diverse scenarios including complex motions, lighting effects, and cultural scenes.



Figure 12: Dragon boat racing scene comparison.

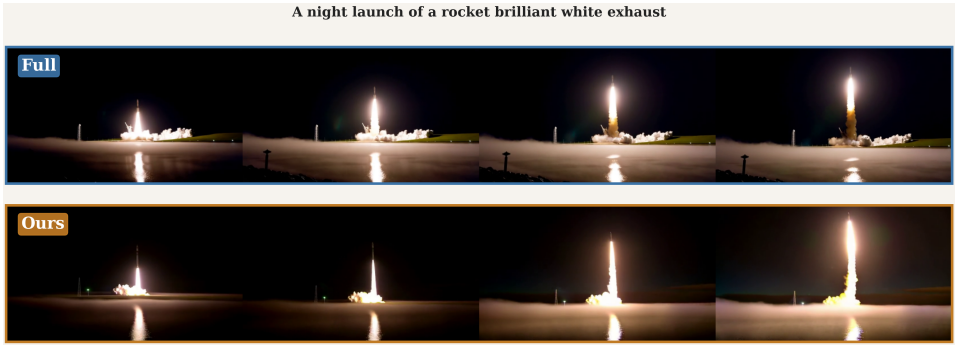

Figure 13: Rocket launch scene comparison.

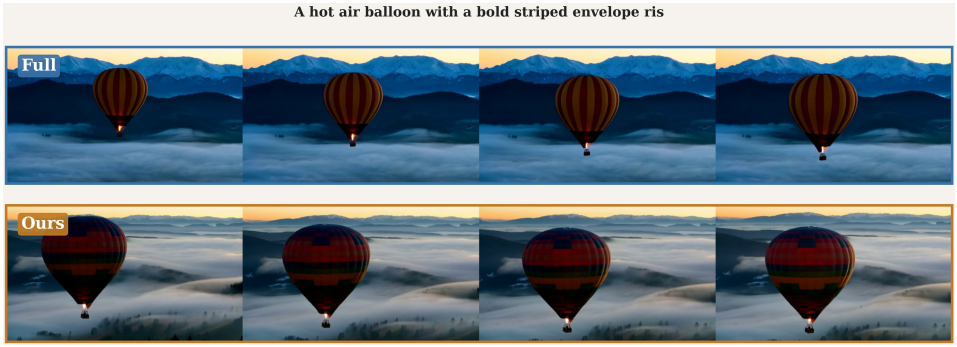

Figure 14: Hot air balloon scene comparison.



Figure 15: Lion dance performance comparison.

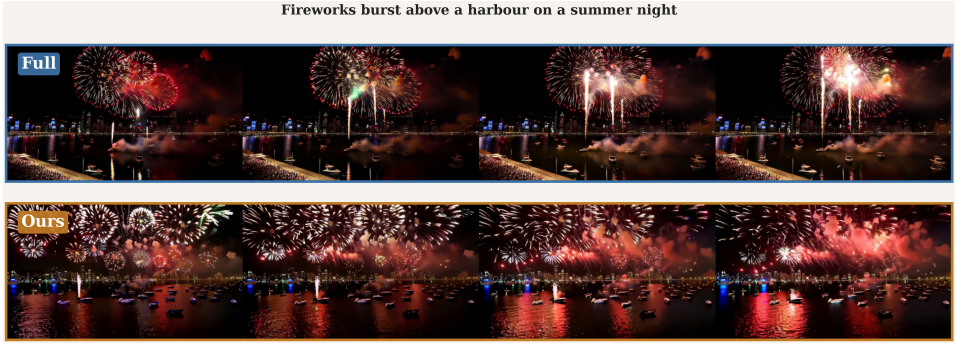

Figure 16: Fireworks display comparison.



Figure 17: Samurai scene comparison.

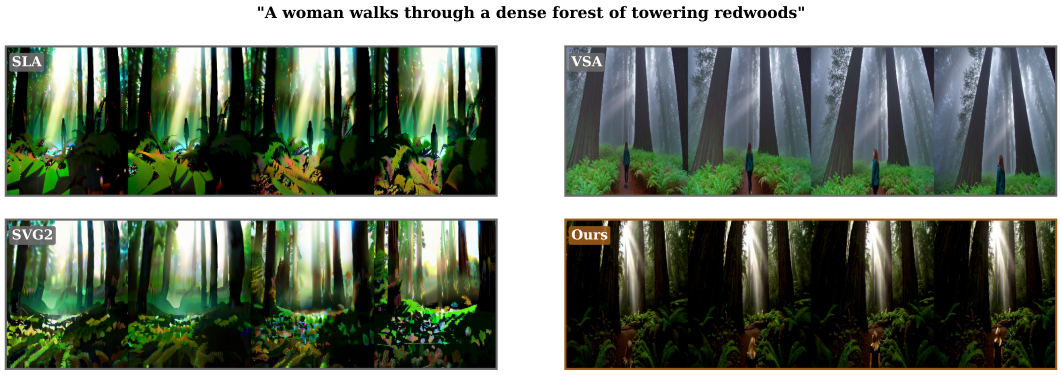

Figure 18: Comprehensive baseline comparison on redwood forest scene. Shows RoPeSLR alongside other sparse attention methods.