DynamicRad: Content-Adaptive Sparse Attention for Long Video Diffusion

Yongji Long[1,2] **Shijun Liang**[3] **Jintao Li**[1,2,*] **Yun Li**[1,2,*]

[1] University of Electronic Science and Technology of China
[2] Shenzhen Institute for Advanced Study, UESTC
[3] Computational Mathematics, Science, & Engineering at Michigan State University (MSU)

[*] Corresponding authors: j.t.li@i4ai.org Yun.Li@ieee.org

Abstract

Leveraging the natural spatiotemporal energy decay in video diffusion offers a path to efficiency, yet relying solely on rigid static masks risks losing critical long-range information in complex dynamics. To address this issue, we propose **DynamicRad**, a sparse-attention framework that constrains adaptive selection using a radial locality prior. DynamicRad introduces a **dual-mode** strategy: *static-ratio* for speed-optimized execution and *dynamic-threshold* for quality-first filtering. To avoid online configuration search, we integrate an offline Bayesian Optimization (BO) pipeline with a **semantic motion router**. The offline-trained router maps prompt embeddings to BO-selected sparsity regimes with a single projection module. Unlike online profiling methods, our offline BO optimizes attention reconstruction error (MSE) on a proxy task and reuses the selected configurations during inference. Dynamic selection still scores current Q/K features and constructs sparse indices. Experiments on HunyuanVideo and Wan2.1-14B demonstrate that DynamicRad achieves a strong efficiency–quality trade-off among block-sparse-attention baselines, obtaining **1.7×–2.5× inference speedups** at **74.3%–86.8% effective block sparsity**. In some long-sequence settings, the dynamic mode matches or slightly improves upon a checkpoint-matched dense baseline on individual automatic quality metrics, while optional Mask-Aware LoRA further improves long-horizon coherence. Code is available at `https://github.com/Adamlong3/DynamicRad`.

1 Introduction

Recent diffusion-based video generators achieve strong visual fidelity and temporal coherence, yet scaling them to long horizons and high resolution remains difficult because self-attention grows quadratically with the spatiotemporal token count [Ho et al., 2020, Song et al., 2021, Rombach et al., 2022, Peebles and Xie, 2023, Ho et al., 2022, Hong et al., 2022, Bar-Tal et al., 2024, Brooks et al., 2024, Vaswani et al., 2017]. Even with optimized kernels, attention-side compute and memory bandwidth remain major bottlenecks [Dao et al., 2022, Dao, 2024, Bolya et al., 2023, Ma et al., 2023].

A natural remedy is sparsity. Prior efficient attention methods mainly fall into two categories: static sparse patterns, which are hardware-friendly but rigid, and dynamic sparsification, which is adaptive but often introduces extra overhead or instability under heterogeneous and long-video settings [Child et al., 2019, Beltagy et al., 2020, Zaheer et al., 2020, Kitaev et al., 2020, Wang et al., 2020, Choromanski et al., 2021, Roy et al., 2021, Tay et al., 2022]. Recent evidence further suggests a spatiotemporal locality structure in video diffusion attention [Li et al., 2025, Zhang et al., 2025b, Ding et al., 2025], where post-softmax attention mass decays with spatial and temporal distance, motivating sparse masks that preserve near-field interactions while pruning long-range

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

pairs [Li et al., 2025, Zhang et al., 2025b]. However, purely static sparsification remains brittle under complex motion, whereas fully dynamic schemes may be costly or unstable at scale.

In this work, we propose **DynamicRad**, a sparse-attention framework for long video diffusion that combines a structured radial candidate set with content-dependent selection. DynamicRad uses a shared block-sparse candidate set and supports two deployment modes: *static-ratio*, which prioritizes throughput by using fixed retention ratios, and *dynamic-threshold*, which uses attention-logit scores to retain candidate pairs. The mode and sparsity configuration are selected from an offline BO-compiled lookup table using a semantic motion router. An optional Mask-Aware LoRA is evaluated separately to compensate for sparsity-induced temporal artifacts. Unlike Radial Attention's content-independent final mask [Li et al., 2025], Dynamic mode selects links from current Q/K features within radial support. Recent methods such as AdaSpa [Xia et al., 2025] also explore training-free dynamic sparsification for long video generation, but rely on online precise search over sparse indices; DynamicRad instead shifts regime-level configuration search offline while Dynamic mode still incurs scoring and indexing costs.

Contributions We summarize our contributions as follows: (1) **Core Framework:** We propose DynamicRad, a unified sparse-attention framework that combines a shared kernel-friendly candidate set with two selection modes, bridging rigid-prior methods such as Radial Attention [Li et al., 2025] and dynamic methods such as SVG [Xi et al., 2025]. We further introduce a multi-dimensional parameterization, including a continuous long-range factor λ, that unifies static efficiency and dynamic flexibility without an online mask-prediction network inside the denoising loop. (2) **Robust Deployment:** We introduce an offline BO-driven configuration pipeline coupled with a lightweight **Semantic Motion Router**. By optimizing **Attention MSE on a physics-based proxy task**, the search process can be completed in minutes, while the router maps prompt embeddings to BO-selected regimes in $<$ 2ms without fragile manual tuning. (3) **Empirical Gains:** Integrating DynamicRad into HunyuanVideo-241f and Wan2.1-14B [Nan et al., 2025, Kong et al., 2025, Wan et al., 2025, Huang et al., 2023], we achieve up to **2.54$\times$ speedup** under high sparsity. We report the **backbone-frozen sparse-attention policy** separately from the optional **Mask-Aware LoRA** adaptation, which further improves temporal coherence under aggressive sparsity.

2 Related Work

Evolution of Video Generation Backbones. Video generation has shifted from U-Net backbones [Ho et al., 2022] to DiT-style architectures [Peebles and Xie, 2023] due to superior scalability. Representative DiT models include Latte [Ma et al., 2025] and CogVideoX [Yang et al., 2024]; autoregressive approaches include VideoPoet [Kondratyuk et al., 2024] and the long-video system StreamingT2V [Henschel et al., 2025].

Sparse Attention & Efficiency Paradigms. Scaling diffusion models to long horizons is hindered by the quadratic $\mathcal{O}(N^2)$ cost of attention [Vaswani et al., 2017, Dao et al., 2022]. Prior works typically bifurcate into *static patterns* that impose fixed locality priors [Child et al., 2019, Li et al., 2025, Zhang et al., 2025b] and *dynamic routing* that adapts to content [Roy et al., 2021, Xi et al., 2025]. While Radial Attention [Li et al., 2025] is hardware-friendly, it lacks adaptivity. Conversely, SVG [Xi et al., 2025] enables dynamic routing but adds online overhead. MOD-DiT [Liu et al., 2026] predicts later masks from early denoising and historical sparsity, whereas our regime router uses the prompt before denoising. BLASST [Yuan et al., 2026] performs threshold pruning for LLM inference rather than video spatiotemporal attention. We treat MA-LoRA as an optional robustness mechanism, not an algorithmic novelty claim.

To address these gaps, recent works have introduced novel structural efficiencies. Dynamic Diffusion Transformer (DyDiT) [Zhao et al., 2024] dynamically adjusts computation in diffusion transformers along timestep and spatial dimensions, while sparse-attention methods such as **PowerAttention (PA)** [Chen et al., 2025] (originally proposed for long-context transformers), **SVG** [Xi et al., 2025], and **Radial Attention** [Li et al., 2025] improve efficiency from different design perspectives.

Learning-based Efficiency. Recent approaches like VORTA [Sun et al., 2025] and VSA [Zhang et al., 2025a] propose learning sparsity policies directly via gating networks. While effective, they require expensive retraining or per-step routing inference. In contrast, DynamicRad achieves

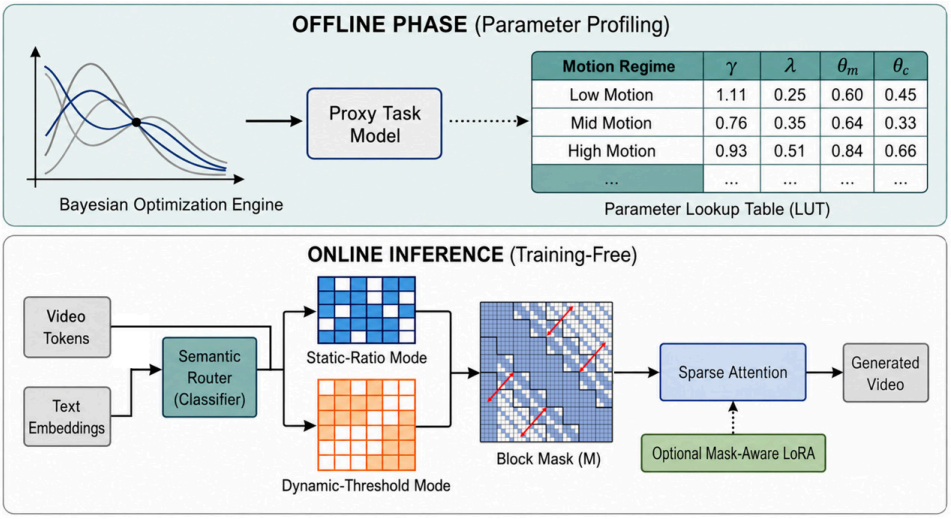


Figure 1: **Overview of DynamicRad.** Offline BO builds a motion-regime lookup table; its displayed values are schematic, while optimized configurations appear in Table 2. At inference, the trained and frozen router selects a configuration from the prompt embedding. Video tokens bypass the router and join the selected Static-ratio or Dynamic-threshold path; only Dynamic scores video Q/K features. The diffusion backbone remains frozen, with optional separate Mask-Aware LoRA.

adaptivity by leveraging physical motion priors and reusing pre-computed text embeddings, avoiding the need for invasive architectural changes. Its router is trained offline and frozen; optional Mask-Aware LoRA is a separate adaptation path.

Deployment Optimization Strategies. Our work integrates several orthogonal efficiency techniques. TeaCache [Liu et al., 2024] and DeepCache [Ma et al., 2023] save compute by reusing cached model outputs or features across denoising steps; we reduce the intra-layer attention cost. In diffusion-based image restoration, Li et al. decouple data-consistency reconstruction from diffusion purification to improve inference efficiency [Li et al., 2026b]. To ensure temporal coherence under aggressive pruning, we employ Parameter-Efficient Fine-Tuning (PEFT) akin to LoRA [Hu et al., 2022], but specifically target projection layers sensitive to sparsity. Finally, we utilize Bayesian Optimization [Snoek et al., 2012] in an offline pipeline to pre-compile a lightweight lookup table, enabling instantaneous adaptation at runtime. More broadly, constrained optimization of competing objectives also appears in policy learning, where comparison-based updates have been used to improve feasible Pareto fronts [Li et al., 2026a].

3 Method

Radial Attention reduces attention cost via a static energy-decay mask [Li et al., 2025], but its fixed sparsity pattern can be brittle under long horizons and complex motion where long-range dependencies become critical. To address these limitations, we propose **DynamicRad**, a sparse-attention framework summarized in Figure 1. The framework combines offline BO-based configuration, prompt-conditioned regime routing, and a shared candidate set with two selection modes: *static-ratio* for throughput-first deployment and *dynamic-threshold* for quality-first deployment. Both modes share the same kernel-friendly block aggregation logic. We additionally introduce an optional Mask-Aware LoRA adaptation path, which is evaluated separately from the backbone-frozen sparse-attention policy.

3.1 Dynamic Neighborhood Partitioning and Parameter Space

We parameterize the attention connectivity used by the Semantic Router through the radial width scale γ, block size B_s, mask threshold θ_m, column density threshold θ_c, and long-range factor λ. The Offline BO engine tunes γ, λ, θ_m, θ_c, and the mode-specific retention ratios or score thresholds,

while B_s is fixed within each run. This formulation allows the radial window width and long-range allocation to adapt to varying motion regimes.

Notation. Let N_f be the number of frames and N_t the number of tokens per frame. The total token count is $S = N_f N_t$. We use $B_s \in \{32, 64, 128\}$ to denote the block size, aligning with hardware granularity. The sequence is padded such that $S' = \lceil S/B_s \rceil \cdot B_s$.

Radial window width via logarithmic grouping. Our construction uses logarithmic distance groups to narrow the radial window as temporal distance increases. For a frame pair (i, j) with temporal distance $t = |i - j|$, we define a logarithmic group index $\mathcal{B}(t)$ and a base token span L_0. We rigorously define the grouping for temporal neighbors ($t \geq 1$):

$$\begin{aligned} \mathcal{B}(t) &= \lfloor \log_2 t \rfloor + 1 \quad (\text{for } t \geq 1), \\ L_0 &= 2^{\lfloor \log_2 N_t \rfloor + 1}. \end{aligned} \tag{1}$$

Here, L_0 represents the implementation-aligned base token span under the flattened latent-token layout. Note that for the self-attention case ($t = 0$), we default to dense attention. The decay length $\mathcal{L}(t)$ is an inverse power-law scaled by γ, so larger γ widens the candidate window:

$$\mathcal{L}(t) = \frac{L_0}{2^{\mathcal{B}(t)}} \cdot \gamma \approx \frac{L_0}{t} \cdot \gamma. \tag{2}$$

The window width $w(i, j)$ is rounded to an integer and lower-bounded by B_s before token-to-block aggregation:

$$w(i,j) = \begin{cases} N_t, & t \leq 1 \text{ (intra \& adjacent)} \\ \max\big(B_s,\ \lfloor \mathcal{L}(t) \rceil\big), & t > 1. \end{cases} \tag{3}$$

Long-range split rule controlled by the long-range factor λ. To capture periodic long-range dependencies (e.g., repetitive motion), we introduce a secondary decay schedule for the "split" stride, controlled by λ:

$$\mathcal{L}_{\text{split}}(t) = \frac{L_0}{2^{\mathcal{B}(t)}} \cdot \lambda. \tag{4}$$

The split stride factor $\mathrm{sf}(t)$ determines the sampling frequency of long-range frames:

$$\mathrm{sf}(t) = \max\left(1, \left\lfloor \frac{B_s}{\mathcal{L}_{\text{split}}(t) + \epsilon} \right\rfloor\right). \tag{5}$$

A frame pair (i, j) is retained only if

$$t \bmod \mathrm{sf}(t) = 0, \tag{6}$$

where $t = |i - j|$ denotes the temporal distance. This rule implies that as t increases, $\mathcal{L}_{\text{split}}(t)$ decreases and the stride $\mathrm{sf}(t)$ increases, thereby sparsifying long-range temporal connections.

Let $u, v \in [0, N_t - 1]$ denote local token indices under the flattened latent-token memory layout within frame i and frame j, respectively. For retained frame pairs, we define the candidate token-pair set as

$$\mathcal{P}(i,j) = \{(u,v) : |u - v| \leq w(i,j)\}, \tag{7}$$

and set $\mathcal{P}(i, j) = \varnothing$ otherwise. Here, $|u - v|$ denotes locality under the flattened latent-token layout before block aggregation. In this way, the radial window $w(i, j)$ controls local token connectivity under the kernel memory layout, while the split rule in Eq. 6 sparsifies distant frame pairs along the temporal dimension.

Density control within blocks. The parameters θ_c and θ_m control the aggregation of token-level links into a block-level mask. For each block, we first mark a column as active if its retained-token ratio exceeds θ_c; the block is then activated if the fraction of active columns exceeds θ_m. Thus, θ_c suppresses noisy columns, while θ_m controls the final block retention for kernel efficiency.

3.2 Two-Mode Intra-Candidate Selection

Given the candidate set $\mathcal{P}(i, j)$ for each inter-frame pair ($t > 0$), DynamicRad selects a subset using one of two mutually exclusive modes. For $t = 0$ (intra-frame) we do not prune and keep dense attention.

Static-ratio mode (static_ratio). Designed for throughput-first deployment, this mode avoids the overhead of content-dependent scoring. We choose a retention ratio based on the decay state:

$$\rho(i,j) = \begin{cases} 1.0, & t \le 1, \\ \rho_1, & t > 1 \text{ and } \mathcal{L}(t) \ge B_s, \\ \rho_2, & t > 1 \text{ and } \mathcal{L}(t) < B_s, \end{cases} \tag{8}$$

and uniformly sample $k = \max(1, \lfloor |\mathcal{P}(i,j)| \cdot \rho(i,j) \rfloor)$ pairs from $\mathcal{P}(i,j)$ using a deterministic seed. The generator is initialized once and is not reset for each frame-pair candidate set, so successive sets receive distinct draws; the seed reproduces the complete mask.

Dynamic-threshold mode (dynamic_threshold). Designed for quality-first deployment, this mode uses attention logits as lightweight importance proxies. For each candidate pair $(u,v) \in \mathcal{P}(i,j)$, we compute a proxy score by averaging pre-softmax logits over a small set of fused heads H_f:

$$s(u,v) = \frac{1}{H_f} \sum_{h=1}^{H_f} \frac{\mathbf{Q}_u^{(h)} \cdot (\mathbf{K}_v^{(h)})^\top}{\sqrt{d_k}}. \tag{9}$$

Score normalization. To make global thresholds more transferable across layers and prompts, we normalize scores within each candidate set $\mathcal{P}(i,j)$:

$$\tilde{s}(u,v) = \frac{s(u,v) - \mu_{ij}}{\sigma_{ij} + \epsilon}, \tag{10}$$

where μ_{ij} and σ_{ij} are the mean and std of $\{s(u',v') : (u',v') \in \mathcal{P}(i,j)\}$. We compute $\tilde{s}(u,v)$ at the token-pair level and then aggregate token links to a block mask via (θ_c, θ_m) to preserve kernel-friendly block sparsity.

Global thresholding with safeguards. We keep pairs whose normalized score exceeds a decay-state-dependent threshold:

$$\tau(i,j) = \begin{cases} -\infty, & t \le 1, \\ \tau_1, & t > 1 \text{ and } \mathcal{L}(t) \ge B_s, \\ \tau_2, & t > 1 \text{ and } \mathcal{L}(t) < B_s. \end{cases} \tag{11}$$

where τ_1 and τ_2 correspond to near_frame_threshold and far_frame_threshold, respectively. In our evaluated settings, pre-norm and scaled dot-product attention yield an empirically stable score scale across layers and denoising steps, making shared thresholds practical; using a small fused-head set ($H_f = 2$) further reduces overhead while preserving rank-order utility for filtering. To avoid fragile per-layer calibration, we use global thresholds (τ_1, τ_2) shared across blocks, together with two safeguards: (i) per-candidate-set score normalization (Eq. 10); (ii) if no pair passes the threshold, we retain the highest-scoring pair as a fallback.

The overall selection procedure is summarized in Algorithm 1, while the detailed definitions of $w(i,j)$, $\mathcal{P}(i,j)$, score normalization, and thresholding are given in Eqs. 3–11.

3.3 Parameter-Efficient Fine-Tuning for Mask-Aware Refinement

Similar to adaptive rank allocation strategies like AdaLoRA [Zhang et al., 2023], our Mask-Aware LoRA selectively refines projections that are most sensitive to the sparsity pattern. While sparse selection reduces computation, aggressive sparsity may impair temporal consistency for long-horizon generation. We adopt a parameter-efficient fine-tuning strategy that adapts only mask-related components while freezing the diffusion backbone.

Where to adapt. Let the attention logits be

$$\mathbf{A} = \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} \in \mathbb{R}^{S' \times S'}, \tag{12}$$

Algorithm 1 DynamicRad: Candidate Construction and Dual-Mode Selection

Require: $\mathbf{Q}, \mathbf{K}, N_f, N_t, B_s$, mode, configuration $\mathbf{x}$
Ensure: Block-sparse mask $\mathbf{M}_b \in \{0,1\}^{S_b \times S_b}$

```
Initialize M_b ← 0
for each frame pair (i, j) do
    t ← |i − j|
    if t = 0 then
        Mark all blocks within the intra-frame region in M_b        ▷ dense intra-frame attention
        continue
    end if
    Compute w(i, j) by Eq. 3 and sf(t) by Eq. 5
    if t mod sf(t) ≠ 0 then
        continue
    end if
    Build P(i, j) by Eq. 7
    if mode = static_ratio then
        Sample k = max(1, ⌊|P(i, j)|ρ(i, j)⌋) pairs
    else
        Compute s̃(u, v) and keep pairs with s̃(u, v) ≥ τ(i, j)
        if no pair is kept then
            Keep the highest-scoring pair
        end if
    end if
    Aggregate selected pairs into a local block mask using (θ_m, θ_c)
    If Dynamic kept pairs but local mask is empty, activate the top-scoring pair's block
    Merge the local block mask into M_b
end for
return M_b
```

and let $\mathbf{M}_b \in \{0,1\}^{S_b \times S_b}$ denote the final block-sparse mask. We expand $\mathbf{M}_b$ to a token-level mask $\widetilde{\mathbf{M}} \in \{0,1\}^{S' \times S'}$ by broadcasting each block to its corresponding token region. The masked attention is computed as

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\Big(\mathbf{A} + \log(\widetilde{\mathbf{M}})\Big)\,\mathbf{V} \tag{13}$$

where the logarithm is applied elementwise with $\log 0 = -\infty$. In DynamicRad, the dynamic threshold mode computes approximate scores (Eq. 9) and applies adaptive thresholds (Eq. 11). We apply LoRA adapters to (i) the selector-related projections and (ii) the Q/K projection layers that shape the score geometry under sparsity, while freezing the rest of the backbone.

3.4 Automated Configuration: Semantic-Aware Motion Routing

Since text prompts inherently encode semantic motion priors (e.g., 'reading' vs. 'running') [Wang et al., 2024, Guo et al., 2024], we leverage the prompt embedding to predict the global motion regime. Optimal sparsity relies heavily on the motion dynamics implied by the text prompt. Naive heuristic rules often fail to capture semantic nuances (e.g., "a fast snail" may imply low absolute motion despite the word "fast"). To address this, we introduce a **Semantic Motion Router** that maps prompt embeddings directly to a BO-selected configuration regime.

Optimization Objective and Efficiency. Since running full video generation for parameter search is computationally prohibitive, we perform BO on the lightweight **Proxy Task** defined in **Appendix C**. The objective is to minimize the **Mean Squared Error (MSE)** between the sparse approximation and the dense proxy attention, subject to a sparsity budget:

$$\begin{aligned}\mathcal{L}(\mathbf{x}) = &\frac{\|\mathbf{A}_{\text{dense}} - \mathbf{A}_{\text{sparse}}(\mathbf{x})\|_F^2}{\|\mathbf{A}_{\text{dense}}\|_F^2} \\ &+ \alpha \cdot \text{ReLU}\big(\mathcal{T} - \mathcal{S}(\mathbf{x})\big)\end{aligned} \tag{14}$$

where $\mathcal{T}$ is the target sparsity (e.g., 80%) and $\mathcal{S}(\mathbf{x})$ is the achieved sparsity under configuration $\mathbf{x}$. Thanks to the low-dimensional search space and the smoothness of the proxy landscape, the TPE optimizer typically **converges within 30 trials**, completing the profiling in < 15 minutes. This ensures that DynamicRad can be rapidly adapted to new resolutions or aspect ratios.

Architecture and Inference. Since modern video diffusion models (e.g., HunyuanVideo) utilize heavy text encoders (e.g., LLM or CLIP) for conditioning, we can reuse the pre-computed text embeddings $\mathbf{e}_{\text{text}}$ to guide sparsity. We append a lightweight 2-layer MLP projection head $\phi(\cdot)$ to predict a motion intensity score:

$$s_{\text{motion}} = \phi(\text{Pool}(\mathbf{e}_{\text{text}})) \in [0, 1]. \tag{15}$$

This score is discretized to select the motion regime $c \in \{\text{Low}, \text{Mid}, \text{High}\}$ (e.g., $s < 0.3 \rightarrow$ Low), which indexes the offline BO lookup table. This router introduces a small runtime overhead. Since $\mathbf{e}_{\text{text}}$ is pre-computed, the router requires a single MLP pass taking $<$ 2ms, which is small relative to the generation latency. At inference time, we encode the prompt, predict a motion regime, retrieve the corresponding sparsity configuration from the lookup table, and run generation with the selected mode.

Training with Optical Flow Proxy. To supervise $\phi(\cdot)$ without expensive human annotation, we employ a physics-grounded proxy task. We construct a dataset $\mathcal{D}_{\text{train}}$ from a subset of OpenVid-1M [Nan et al., 2025]. For each video-text pair, we use RAFT [Teed and Deng, 2020] to estimate optical flow between adjacent frames and take its average magnitude as the motion supervision target y_{flow}. The projection head is trained to minimize the regression loss $\mathcal{L} = \|s_{\text{motion}} - \text{Norm}(y_{\text{flow}})\|_2^2$. This aligns the semantic understanding of the router with the physical motion priors used in our offline BO optimization.

4 Experiments

4.1 Experimental Setup

Implementation & Hardware. We implement DynamicRad using **FlashInfer** kernels ($B_s = 32$) on **NVIDIA H100 GPUs**. Standard inference uses a single H100, while HunyuanVideo-241f (>200k tokens) employs an **8×H100 cluster** (sequence parallelism) to prevent OOM. The **Semantic Motion Router** is frozen during inference; its parameter search follows the offline BO pipeline (Sec. 3.4). Unless otherwise specified, the main comparisons marked with † report the **backbone-frozen DynamicRad sparse-attention policy** with an offline-trained, frozen router and without Mask-Aware LoRA. Mask-Aware LoRA ($r = 16$) is fine-tuned for 1k steps on a subset of OpenVid-1M using the standard diffusion denoising loss under the selected sparse mask, while keeping the backbone frozen, and is reported as an optional adaptation setting. Detailed hyperparameters, overhead quantification, and proxy-task analysis are provided in the supplementary material.

Baselines & Protocols. We compare DynamicRad against state-of-the-art sparse-attention baselines using a common block-sparse backend. Specifically, we evaluate **PowerAttention (PA)** [Chen et al., 2025], **SVG** [Xi et al., 2025], and **Radial Attention** [Li et al., 2025], which are reported in Table 1, and our reproduction of **AdaSpa** [Xia et al., 2025] in Appendix E. *Exclusion of other methods:* We discuss MOD-DiT [Liu et al., 2026] and BSA [Zhan et al., 2025], but do not include them in the controlled timing table, which evaluates the implemented baselines with a common block-sparse backend. We also exclude orthogonal kernel-level sparse-attention systems, such as LiteAttention [Shmilovich et al., 2025] and BLASST [Yuan et al., 2026], because their implementation assumptions and target deployment settings differ from our block-sparse radial attention framework.

Following [Li et al., 2025], long-context sparse settings (241f/145f) use HunyuanVideo and Wan2.1-14B checkpoints fine-tuned on **OpenVid-1M** [Nan et al., 2025], while standard-length settings use original checkpoints; unless Mask-Aware LoRA is enabled, DynamicRad remains an inference-time sparse-attention policy. The Original-LX rows instead extrapolate native checkpoints without length adaptation. Full-FT Dense uses the same target-length-adapted checkpoint as the matched sparse variants and serves as their quality reference. The Semantic Motion Router was trained on 20k OpenVid-1M video-text pairs using RAFT-derived optical-flow labels (taking $<$ 1h on a single H100). For evaluation, we report **VBench** scores [Huang et al., 2023] for subject consistency and

Table 1: **Quality–Efficiency Comparison.** All videos are 768p (1280×768); HunyuanVideo-241f uses 8×H100, and other settings use one H100. Speedups use Original (standard length) or Original-LX (long length). Full-FT Dense is a checkpoint-matched quality reference; its latency and speedup are — because it is outside this efficiency comparison. †: DynamicRad without Mask-Aware LoRA, with an offline-trained frozen router; ‡: long-context SVG reproduction. Sparsity is the fraction of padded dense-grid blocks absent from the final BSR execution index. S.C./A.Q./I.Q. denote VBench subject consistency, aesthetic quality, and imaging quality. **Bold**: overall best; underline: best baseline, or best DynamicRad when a baseline leads.

Model	#Frames	Method	Quality ↑				Efficiency		
			VisionReward	S.C.	A.Q.	I.Q.	Sparsity (%)	Latency (s) ↓	Speedup
HunyuanVideo	120	Original	0.141	0.962	0.645	0.675	0.0	1705	1.00
		PA	0.138	0.956	0.637	0.662	70.5	1024	1.67
		SVG	0.144	**0.965**	0.651	0.679	78.2	872	1.96
		Radial Attention	0.140	0.961	0.646	0.673	76.6	884	1.93
		Ours (Static)†	0.138	0.959	0.642	0.668	**84.8**	**861**	**1.98**
		Ours (Dynamic)†	**0.146**	0.961	**0.653**	**0.681**	81.6	917	1.86
	241	Original-LX	0.116	0.948	0.592	0.597	0.0	824	1.00
		Full-FT Dense	0.122	0.949	0.619	0.644	0.0	—	—
		PA	0.108	0.960	0.611	0.649	80.4	346	2.38
		SVG‡	0.121	0.957	0.614	0.651	80.2	408	2.02
		Radial Attention	0.125	0.961	0.626	0.659	80.6	351	2.35
		Ours (Static)†	0.118	0.951	0.612	0.668	**86.8**	**324**	**2.54**
		Ours (Dynamic)†	**0.131**	**0.969**	**0.659**	**0.687**	83.6	378	2.18
		Ours (Routed)†	0.127	0.965	0.642	0.676	84.7	342	2.41
Wan2.1-14B	72	Original	**0.136**	0.958	0.620	0.650	0.0	1680	1.00
		PA	0.126	0.944	0.594	0.620	66.3	1040	1.62
		SVG	0.116	0.938	0.580	0.605	70.2	990	1.70
		Radial Attention	0.129	0.952	0.610	0.641	70.4	945	1.78
		Ours (Static)†	0.127	0.949	0.606	0.637	**78.8**	**875**	**1.92**
		Ours (Dynamic)†	0.134	**0.959**	**0.621**	**0.656**	74.3	980	1.71
	145	Original-LX	0.098	0.921	0.573	0.585	0.0	5160	1.00
		Full-FT Dense	0.146	0.958	0.586	0.673	0.0	—	—
		PA	0.132	0.963	0.592	0.642	73.4	2550	2.02
		SVG‡	0.133	0.965	0.596	0.648	73.8	3035	1.70
		Radial Attention	0.142	0.972	0.602	0.661	73.3	2560	2.02
		Ours (Static)†	0.137	0.969	0.606	0.647	**81.6**	**2289**	**2.25**
		Ours (Dynamic)†	**0.148**	**0.979**	**0.621**	**0.686**	77.0	2850	1.81
		Ours (Routed)†	0.144	0.974	0.613	0.670	78.6	2490	2.07

VisionReward [Xu et al., 2026] for human-aligned aesthetic quality. VBench Total and Dynamic Degree appear in Appendix G.

4.2 Main Results

Efficiency–quality trade-off. Table 1 shows that DynamicRad provides a strong efficiency–quality trade-off among the evaluated sparse-attention baselines. The *static-ratio* mode serves as a throughput-first operating point. On HunyuanVideo-241f, it achieves a **2.54× speedup** (324s), surpassing Radial Attention (351s). Similarly, on Wan2.1-14B (145f), it achieves a **2.25× speedup** (vs. Radial Attention's 2.02×), reducing inference time compared to the dense baseline. The Routed operating point takes 342/2490s at 241f/145f versus Radial Attention's 351/2560s and has slightly higher reported quality, but the sparsity budgets differ; this is a trade-off comparison rather than matched-budget dominance.

Quality preservation. The *dynamic-threshold* mode preserves quality under high sparsity. On HunyuanVideo-241f, it improves over checkpoint-matched Full-FT Dense on VisionReward (0.131 vs. 0.122). We hypothesize that content-aware filtering removes low-utility long-range interactions in long sequences, although this effect may depend on the prompt and model setting. On Wan-145f, VisionReward is 0.148 versus 0.146 for matched dense. VBench Total differs by only +0.14/+0.17 points (Appendix G), supporting quality preservation under the matched evaluation protocol.

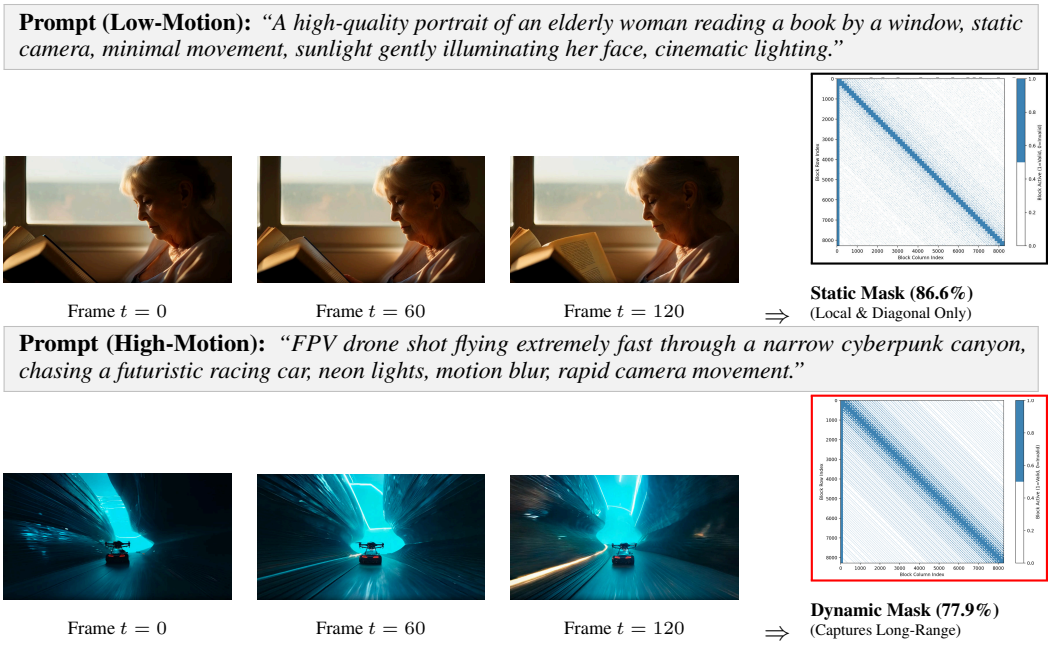


Figure 2: **Regime-aware adaptivity.** For low-motion prompts, the router selects *static-ratio*, yielding a highly sparse near-diagonal mask. For high-motion prompts, it switches to *dynamic-threshold*, preserving off-diagonal long-range connections useful for fast motion.

Robustness Across Architectures. The gains are consistent across HunyuanVideo and Wan2.1-14B. On Wan2.1-14B (72f), static mode maintains a **1.92× speedup** with **78.8% sparsity**, while dynamic mode matches dense temporal consistency (VBench S.C. 0.959 vs. 0.958), suggesting transfer across different tokenization strategies. An A800 run confirms the same acceleration trend (Appendix D).

4.3 Ablation Study

Configurable Parameters & Dual-Mode Selection. The introduction of configurable knobs (λ, B_s, θ) is crucial for stability across resolutions. On top of this, the selection mode serves as the primary trade-off lever: **Static-ratio mode** acts as an efficiency-first option. Compared to a naive configurable baseline, our static optimization (Dual-mode pruning) reduces latency to **338s**. **Dynamic-threshold mode** utilizes attention scores (Eq. 9) to filter tokens. Although it introduces a slight overhead (385s in this ablation), it recovers fine-grained details (VisionReward stabilizes around 0.119) relative to the selector-only Static row, which is not the deployed Static operating point.

Regime-Aware Adaptivity (Offline Predictor). The **Offline BO Predictor** (Sec. 3.4) is an important component of DynamicRad. As shown in **Table 3**, the predictor identifies BO-selected sparsity regimes for the specific motion context, effectively reducing latency further (Static mode: $338s \rightarrow \mathbf{324s}$) while improving quality (VisionReward $0.117 \rightarrow 0.122$). **Table 2** and **Figure 2** illustrate this trend: the system assigns aggressive static masking ($\gamma = 2.0$, **86.6%** sparsity) to low-motion prompts (e.g., reading) while automatically switching to dynamic selection with denser long-range connections for high-motion inputs.

Mask-Aware LoRA. Finally, integrating **Mask-Aware LoRA** (Sec. 3.3) further improves temporal consistency (VBench S.C. reaches **0.977**). By refining the logit distribution to accommodate the sparse structure, MA-LoRA mitigates aggressive pruning artifacts with a modest latency increase from 378s to 389s.

Table 2: **BO-Optimized Configurations.** Low motion permits 86.6% sparsity; high motion uses larger λ for denser long-range links. Near/far values are retention ratios ρ (Static) or thresholds τ (Dynamic). Green and red shading mark speed- and quality-oriented configurations.

Motion	Mode	Width (γ)	Long (λ)	Mask (θ_m)	Column (θ_c)	Near (ρ_1/τ_1)	Far (ρ_2/τ_2)	Best Sparsity	Rationale
Low	Static	2.0	0.3	0.75	0.20	0.25	0.55	**86.6%**	Low retention; sparse long-range.
	Dynamic	1.3	0.4	0.70	0.45	-1.2	2.4	82.7%	Balanced profile.
Mid	Static	1.3	0.7	0.60	0.15	0.75	0.55	84.0%	Target satisfied.
	Dynamic	1.4	0.7	0.70	0.45	-1.5	2.0	80.6%	Target satisfied.
High	Static	1.6	0.8	0.65	0.15	0.55	0.50	81.7%	Denser local budget.
	Dynamic	1.6	0.9	0.55	0.40	-2.0	2.4	**77.9%**	Higher λ; loose thresholds.

Table 3: **Component ablations on HunyuanVideo-241f at 768p (8×H100).** The backbone-frozen sparse policy and optional Mask-Aware LoRA are evaluated on a fixed prompt–seed subset; quality scores may therefore differ slightly from Table 1.

Variant (HunyuanVideo, 241f, 768p, **8×H100**)	Sparsity↑	Latency(s)↓	VisionReward↑	VBench S.C.↑
Radial Attention + Configurable params	81.3%	351	0.123	0.956
+ Dual-mode pruning (Ours, Static mode)	85.5%	338	0.117	0.946
+ Dual-mode pruning (Ours, Dynamic mode)	82.3%	385	0.119	0.948
+ Offline predictor (auto config, Static mode)	**86.8%**	**324**	0.122	0.956
+ Offline predictor (auto config, Dynamic mode)	83.6%	378	0.127	0.963
+ MA-LoRA (Mask-Aware LoRA, applied on Dynamic mode)	83.6%	389	**0.143**	**0.977**

Micro-Benchmark of Routing Overhead. On H100/Wan-72f, the deployed Dynamic path measures **1.71 ms** per invocation for QK scoring, thresholding, block aggregation, index construction, and planning; the router runs once per video. Complete generation takes 980 ± 9s versus 1680 ± 15s for dense attention ($1.71\times$ speedup), including all generation and host overheads. Appendix D gives the component protocol and A800 result.

5 Conclusion

We presented **DynamicRad**, a sparse-attention framework that combines a kernel-friendly radial candidate set with content-adaptive selection for long video diffusion. By implementing a dual-mode strategy via an offline BO-driven router, DynamicRad achieves **1.7×–2.5× speedups** with **74.3%–86.8% effective block sparsity** on HunyuanVideo and Wan2.1-14B. As an optional adaptation path, lightweight Mask-Aware LoRA further preserves temporal coherence under aggressive pruning.

Limitations and Future Work. DynamicRad still has several boundary conditions. (1) **Extreme Fast Motion:** Under highly dynamic prompts (e.g., FPV drone racing), minor artifacts may still appear because localized block-sparse patterns cannot fully capture highly dispersed attention. (2) **Ambiguous or Multi-Phase Prompts:** The Semantic Motion Router predicts a global motion regime from text; for ambiguous prompts or videos with temporal phase shifts, a single fixed regime may be suboptimal, although dynamic fallbacks mitigate this issue. (3) **Hardware Generalization:** Because the method relies on specific block sizes (B_s), non-server GPUs may require local tuning and LUT recalibration. Future work will study frame-level dynamic routing and extend the regime-aware design to image-to-video generation.

Acknowledgments and Disclosure of Funding

The project was supported by the National Natural Science Foundation of China (Grant No. W2431044).

The authors have no competing interests to disclose.

References

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711312. doi:10.1145/3680528.3687614. URL https://doi.org/10.1145/3680528.3687614.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. URL https://arxiv.org/abs/2004.05150.

James S. Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, volume 24, 2011. URL https://papers.nips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations (ICLR)*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, et al. Video generation models as world simulators, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

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

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019. URL https://arxiv.org/abs/1904.10509.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, et al. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Hangliang Ding, Dacheng Li, Runlong Su, Peiyuan Zhang, Zhijie Deng, Ion Stoica, and Hao Zhang. Efficient-vDiT: Efficient video diffusion transformers with attention tile, 2025. URL https://arxiv.org/abs/2502.06155.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning, 2024. URL https://arxiv.org/abs/2307.04725.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. StreamingT2V: Consistent, dynamic, and extendable long video generation from text, 2025. URL https://arxiv.org/abs/2403.14773.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, et al. Imagen Video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. CogVideo: Large-scale pretraining for text-to-video generation via transformers, 2022. URL https://arxiv.org/abs/2205.15868.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, et al. VBench: Comprehensive benchmark suite for video generative models, 2023. URL https://arxiv.org/abs/2311.17982.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations (ICLR)*, 2020.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, et al. VideoPoet: A large language model for zero-shot video generation, 2024. URL https://arxiv.org/abs/2312.14125.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, et al. HunyuanVideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

Jintao Li, Maowen Tang, Yongji Long, Weixuan Liu, Yanlang Zheng, Sicheng He, Ao-Jin Li, Shui Yu, and Yun Li. Learning from comparison: Constrained projection policy optimization for pareto-front improvement. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 70309–70335, 2026a. URL https://proceedings.mlr.press/v306/li26ga.html.

Xiang Li, Soo Min Kwon, Shijun Liang, Ismail Alkhouri, Saiprasad Ravishankar, and Qing Qu. DCDP: Decoupled data consistency via diffusion purification for inverse problems. In *2026 IEEE International Conference on Computational Photography (ICCP)*, 2026b. doi:10.1109/ICCP69532.2026.11668875. URL https://doi.org/10.1109/ICCP69532.2026.11668875.

Xingyang Li, Muyang Li, Tianle Cai, Haocheng Xi, et al. Radial attention: $\mathcal{O}(n \log n)$ sparse attention with energy decay for long video generation. *arXiv preprint arXiv:2506.19852*, 2025.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model, 2024. URL https://arxiv.org/abs/2411.19108.

Yuxi Liu, Yipeng Hu, Zekun Zhang, Kunze Jiang, and Kun Yuan. Mixture of distributions matters: Dynamic sparse attention for efficient video diffusion transformers, 2026. URL https://arxiv.org/abs/2601.11641.

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, 2025. URL https://arxiv.org/abs/2401.03048.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. DeepCache: Accelerating diffusion models for free, 2023. URL https://arxiv.org/abs/2312.00858.

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, 2025. URL https://arxiv.org/abs/2407.02371.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021. doi:10.1162/tacl_a_00353. URL https://aclanthology.org/2021.tacl-1.4/.

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.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Wenhao Sun, Rong-Cheng Tu, Yifu Ding, Zhao Jin, Jingyi Liao, Shunyu Liu, and Dacheng Tao. VORTA: Efficient video diffusion via routing sparse attention, 2025. URL https://arxiv.org/abs/2505.18809.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey, 2022. URL https://arxiv.org/abs/2009.06732.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision (ECCV)*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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

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

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation, 2024. URL https://arxiv.org/abs/2312.03641.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, et al. Sparse VideoGen: Accelerating video diffusion transformers with spatial-temporal sparsity. *arXiv preprint arXiv:2502.01776*, 2025.

Yifei Xia, Suhan Ling, Fangcheng Fu, Yujie Wang, Huixia Li, Xuefeng Xiao, and Bin Cui. Training-free and adaptive sparse attention for efficient long video generation, 2025. URL https://arxiv.org/abs/2502.21079.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, et al. VisionReward: Fine-grained multi-dimensional human preference learning for image and video generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(13): 11269–11277, 2026. doi:10.1609/aaai.v40i13.38107.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Jiayi Yuan, Cameron Shinn, Kai Xu, Jingze Cui, George Klimiashvili, Guangxuan Xiao, et al. BLASST: Dynamic blocked attention sparsity via softmax thresholding, 2026. URL https://arxiv.org/abs/2512.12087.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, et al. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Chenlu Zhan, Wen Li, Chuyu Shen, Jun Zhang, Suhui Wu, and Hao Zhang. Bidirectional sparse attention for faster video diffusion training, 2025. URL https://arxiv.org/abs/2509.01085.

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. *arXiv preprint arXiv:2505.13389*, 2025a.

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

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning, 2023. URL https://arxiv.org/abs/2303.10512.

Wangbo Zhao, Yizeng Han, Jiasheng Tang, Kai Wang, Yibing Song, Gao Huang, Fan Wang, and Yang You. Dynamic diffusion transformer, 2024. URL https://arxiv.org/abs/2410.03456.

A Illustrative Complexity Scaling

This section compares dense attention with an idealized radial-decay model to illustrate the scaling of the locality prior. Let N_f denote the number of frames, N_t the number of tokens per frame, and D the feature dimension. The total sequence length is $S = N_f N_t$.

A.1 Standard Dense Attention

The standard self-attention mechanism computes the similarity between all pairs of tokens. The computational cost consists of two parts: the matrix multiplication $\mathbf{QK}^\top$ and the attention-value aggregation $\mathbf{AV}$.

$$\mathcal{C}_{\text{dense}} = \mathcal{O}(S^2 D) = \mathcal{O}(N_f^2 N_t^2 D). \tag{16}$$

The memory complexity for storing the attention map is $\mathcal{O}(S^2)$, which is quadratic with respect to the frame count N_f.

A.2 Idealized Radial-Decay Model

For this illustration, let the inter-frame window be $\widetilde{w}(t) = L_0\gamma/t$ for $t \geq 1$, before applying the minimum width B_s, the long-range split rule, either mode's intra-candidate selection, or block aggregation. There are at most two frames at each temporal distance from frame i, so for $N_f \geq 2$ the window-sum proxy satisfies

$$\widetilde{\Omega}_i = \sum_{\substack{j=0 \\ j\neq i}}^{N_f-1} \widetilde{w}(|i-j|) \leq 2L_0\gamma \sum_{t=1}^{N_f-1} \frac{1}{t} \leq 2L_0\gamma\big(1 + \ln(N_f - 1)\big). \tag{17}$$

Assuming attention arithmetic is proportional to this idealized link count, and treating γ as constant, $L_0 = \Theta(N_t)$ gives

$$\widetilde{\mathcal{C}}_{\text{radial}} = \mathcal{O}(S\widetilde{\Omega}_{\text{avg}} D) = \mathcal{O}(N_f N_t^2 \ln N_f D). \tag{18}$$

Relative to dense attention, this idealized arithmetic estimate scales as $\mathcal{O}(\ln N_f / N_f)$ in the frame count. It is an illustration of the radial prior, not a complexity bound for the deployed DynamicRad algorithm: the minimum block width, split retention, mode-specific selection, block aggregation, and mask-construction overhead determine its actual cost.

B Implementation Details: Efficiency and Robustness

A common concern for dynamic sparse attention is the trade-off between implementation overhead and stability. In this section, we detail how DynamicRad minimizes computational costs via proxy scoring and ensures robustness via fallback mechanisms.

Head Fusion Proxy (Method). Instead of computing scores across all heads (e.g., $H = 32$), we hypothesize that attention structures are shared. We compute filtering logits using only the first $H_f = 2$ heads:

$$S_{ij}^{\text{proxy}} = \frac{1}{H_f} \sum_{h=1}^{H_f} \frac{Q_{i,h} K_{j,h}^\top}{\sqrt{d}}. \tag{19}$$

Theoretical Overhead Analysis. The following summarizes the theoretical cost reductions; a measured H100/Wan-72f latency breakdown is reported in Appendix D:

- **Selection Cost Reduction:** By using $H_f = 2$ out of $H = 32$ heads, the logit computation cost is reduced by a factor of **16**×. The subsequent sorting and thresholding are performed on this reduced subset, ensuring the selection overhead remains a small fraction of the full $\mathcal{O}(N^2)$ attention.
- **Router Efficiency:** The Semantic Motion Router uses a lightweight MLP on pre-computed embeddings. Its complexity is $\mathcal{O}(1)$ relative to the video length N_f, incurring negligible cost compared to the quadratic attention mechanism.

Robustness and Fallback Mechanisms. For ambiguous prompts where the Semantic Router predicts low confidence (or classifies incorrectly), DynamicRad employs two safeguards to prevent quality collapse:

1. **Score-Based Fallback:** In the *dynamic-threshold* mode, if the threshold rejects every pair in a nonempty frame-pair candidate set, we retain its highest-scoring pair. If selected pairs are nonempty but no block survives aggregation, we activate the block containing the highest-scoring candidate pair. This prevents the final block mask for the retained frame pair from being empty.
2. **Conservative Default:** If the router output is out-of-distribution (e.g., gibberish prompts), the system defaults to the *Static-ratio (Mid-Motion)* profile, providing a safe baseline performance similar to standard sparse attention.

C Offline Bayesian Optimization Details

To ensure DynamicRad adapts robustly to varying video dynamics without incurring online tuning costs, we utilize a rigorous offline Bayesian Optimization (BO) strategy. This section details the proxy task formulation, clarifies its scope, and analyzes the optimization convergence.

C.1 Detailed Hyperparameter Configuration

To support reproducibility, Table 4 summarizes the fixed system parameters and the search ranges explored by the Bayesian optimizer. Structural parameters are fixed to match hardware-aware kernel constraints, whereas sparsity-related filtering parameters are optimized offline within the ranges listed below.

Table 4: **Hyperparameter Settings and Optimization Bounds.** System parameters are fixed within each run rather than optimized by BO. For mode-specific parameters, we report the **Search Range** $[min, max]$ explored by the Bayesian Optimizer.

I. Fixed System Parameters			
Parameter	**Value**	**Description**	
Block Size (B_s)	128/64/32	Fixed granularity for FlashInfer/Sage kernels.	
Fused Heads (H_f)	2	Fixed proxy for fast dynamic scoring.	
II. Adaptive Mode Parameters (Search Ranges)			
Mode	**Parameter**	**Optimization Range (via BO)**	**Description**
Common	Width Scale (γ)	$[1.0, 3.0]$	Controls radial window width.
	Long Factor (λ)	$[0.1, 1.0]$	Controls long-range split density.
	Col Density Thresh (θ_c)	$[0.1, 1.0]$	Minimal density to consider a column "active".
	Mask Threshold (θ_m)	$[0.1, 1.0]$	Range for block activation threshold.
Static-ratio	topk_ratio1	$[0.1, 1.0]$	For near-field blocks.
	topk_ratio2	$[0.1, 0.8]$	For far-field blocks.
Dynamic-threshold	near_frame_thresh	$[-10.0, 5.0]$	Threshold on $\tilde{s}$ for $t > 1$, $\mathcal{L}(t) \geq B_s$.
	far_frame_thresh	$[-5.0, 8.0]$	Threshold on $\tilde{s}$ for $t > 1$, $\mathcal{L}(t) < B_s$.

C.2 Mathematical Formulation of the Proxy Task

Running full diffusion generation loops for hyperparameter search is computationally prohibitive. Instead, we construct a lightweight **proxy task** that simulates the feature drift inherent in video attention.

Let $\mathbf{h}_t \in \mathbb{R}^D$ denote the hidden state (Key/Query feature) of a token at frame t. We model the temporal evolution of features as an auto-regressive process governed by a drift rate $\delta \in [0, 1]$:

$$\mathbf{h}_t = \sqrt{1 - \delta^2} \cdot \mathbf{h}_{t-1} + \delta \cdot \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{20}$$

where $\boldsymbol{\epsilon}_t$ represents new semantic information entering the frame.

- **Low Motion ($\delta = 0.02$):** Features remain highly correlated ($\mathbf{h}_t \approx \mathbf{h}_{t-1}$), mimicking static backgrounds.

- **Mid Motion ($\delta = 0.15$):** Features exhibit moderate drift, representing standard character movements or slow camera pans.
- **High Motion ($\delta = 0.40$):** Features drift rapidly, simulating fast-moving objects or rapid scene changes.

Alignment with Semantic Router. RAFT-derived optical-flow magnitude defines the real-video motion regimes that supervise the router (Section 3.4). The simulator parameter δ is a dimensionless feature-drift magnitude, not a conversion from pixel-space optical flow. Values 0.02, 0.15, and 0.40 are representative low-, mid-, and high-drift stress levels ordered consistently with those regimes.

The proxy attention matrix $\mathbf{A}_{\text{proxy}}$ is then computed as $\mathbf{A}_{ij} \propto \exp(\mathbf{h}_i^\top \mathbf{h}_j / \sqrt{D})$.

C.3 Scope of the Proxy-Model Analysis

The AR proxy is a controlled test of how sparsity settings respond to feature drift, not a formal model of every video-diffusion attention pattern. A locality prior can discard useful long-range links. Dynamic-threshold mode scores only pairs inside the radial candidate set $\mathcal{P}(i, j)$, so it cannot recover links excluded by that set; within it, score-based selection can retain high-scoring pairs that static-ratio sampling may miss. We do not claim a deterministic error advantage for Dynamic mode.

C.4 Optimization Objective

We search a mode-specific parameter vector, with the block size B_s fixed within each profiling run:

$$\begin{aligned} \mathbf{x}_{\text{static}} &= [\gamma, \lambda, \theta_m, \theta_c, \rho_1, \rho_2], \\ \mathbf{x}_{\text{dynamic}} &= [\gamma, \lambda, \theta_m, \theta_c, \tau_1, \tau_2]. \end{aligned} \tag{21}$$

For either mode, $\mathbf{x}$ denotes its corresponding vector. The objective balances reconstruction fidelity (approximating the dense attention matrix) and computational efficiency (sparsity target):

$$\mathcal{L}(\mathbf{x}) = \underbrace{\frac{||\mathbf{A}_{\text{dense}} - \mathbf{A}_{\text{sparse}}(\mathbf{x})||_F^2}{||\mathbf{A}_{\text{dense}}||_F^2}}_{\text{Normalized MSE}} + \alpha \cdot \underbrace{\text{ReLU}(\mathcal{T} - \mathcal{S}(\mathbf{x}))}_{\text{Sparsity Penalty}} \tag{22}$$

where $\mathcal{S}(\mathbf{x})$ is the sparsity ratio, $\mathcal{T} = 0.80$ is the target sparsity, and $\alpha = 10.0$ is a penalty coefficient.

C.5 Optimization Algorithm and Efficiency

We employ the Tree-structured Parzen Estimator (TPE) algorithm [Bergstra et al., 2011] to explore the non-differentiable search space efficiently. The complete offline profiling procedure is summarized in Algorithm 2.

Efficiency Analysis. While our proxy task requires computing dense attention baselines for loss calculation, it remains significantly faster than standard hyperparameter tuning on video diffusion models. Replacing proxy evaluations with full video generations would require repeated generation runs for each candidate configuration and motion regime. Our proxy-based optimization completes the reported calibration in approximately **15 minutes**, allowing adaptation to new resolutions or aspect ratios without iterative video sampling.

C.6 Convergence Analysis and Qualitative Validation

Convergence Behaviors (Figure 3). We first analyze the optimization trajectories to validate the necessity of our dual-mode design. As shown in Figure 3:

- **Static Mode Limitation:** For High Motion scenes (red line, Left), the loss stagnates near 1.0 over the tested parameter range. This suggests that the tested static masks do not capture this proxy's rapid feature drift as well as Dynamic mode.
- **Dynamic Mode Recovery:** In contrast, the Dynamic mode (red line, Right) drives the loss down to $\sim$0.30. By retaining high-scoring off-diagonal pairs within radial support, it better captures the long-range dependencies represented in this proxy.

Algorithm 2 Offline Parameter Profiling via TPE

Require: Proxy Simulator $\mathcal{P}(\delta)$, Search Space Ω, Trials $N = 30$.
Ensure: Optimal configs $\{\mathbf{x}^*_{\text{low}}, \mathbf{x}^*_{\text{mid}}, \mathbf{x}^*_{\text{high}}\}$.

for regime $r \in \{\text{Low}, \text{Mid}, \text{High}\}$ defined in Sec. 3.4 **do**
  Set drift rate δ_r (0.02, 0.15, or 0.40).
  Initialize trial history $\mathcal{H} \leftarrow \emptyset$.
  for $i = 1$ to N **do**
    Sample $\mathbf{x}_i$ from Ω using TPE acquisition function based on $\mathcal{H}$.
    Generate features $\mathbf{Q}, \mathbf{K}$ using $\mathcal{P}(\delta_r)$.
    Compute Sparse Mask $\mathbf{M}(\mathbf{x}_i)$.
    Calculate Loss $\ell_i = \mathcal{L}(\mathbf{x}_i)$ via Eq. (22).
    Update history $\mathcal{H} \leftarrow \mathcal{H} \cup \{(\mathbf{x}_i, \ell_i)\}$.
  end for
  $\mathbf{x}^*_r \leftarrow \arg\min_{\mathbf{x} \in \mathcal{H}} \mathcal{L}(\mathbf{x})$
end for
return Lookup Table $\mathcal{T}$ ▷ Indexed by Semantic Router outputs

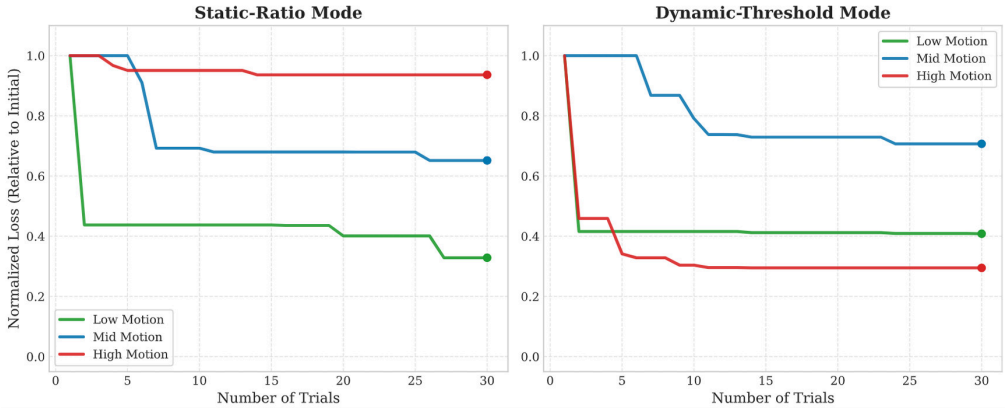


Figure 3: **BO Convergence Analysis.** The optimization process (simulated on proxy tasks) typically converges within 30 trials. **Left:** Static mode struggles with High Motion (red line stays high), necessitating a mode switch. **Right:** Dynamic mode successfully minimizes reconstruction error for High Motion scenes, validating its capability to capture complex dynamics.

Validation of Proxy Alignment (Figure 4). The visual ablation illustrates the association between lower proxy MSE and better generations; a matched-sparsity quantitative analysis is reported in Appendix F.

- **High MSE → Artifacts:** As shown in Figure 4(a), the high reconstruction error in static mode is associated with visual artifacts (flickering neon rings) and geometric distortion.
- **Low MSE → Fidelity:** Conversely, Figure 4(b) demonstrates that when the BO pipeline minimizes the Proxy MSE, the visual fidelity is improved.

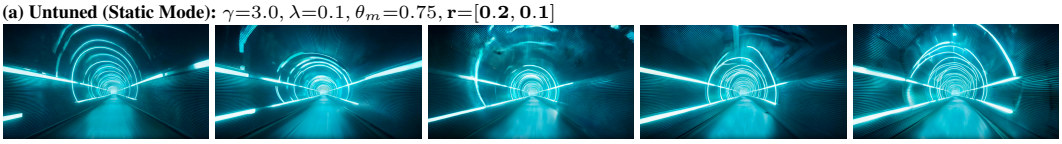
Together, the qualitative and quantitative results support proxy MSE as a configuration-ranking signal under the tested prompts and sparsity range, rather than a universal predictor of video quality.

D Measured Dynamic-Mode Overhead

We instrument the deployed Wan2.1-14B-72f pipeline on one H100 under the Table 1 configuration. After 20 warm-up iterations, component latencies are measured over 200 iterations using synchronized CUDA events for GPU stages and synchronized wall-clock timing for host-side index construction and planning. Each Dynamic invocation builds its mask, CSR/BSR indices, and execution plan from current Q/K activations; no content-dependent result is reused across layers or denoising steps. The router runs once per video from its prompt embedding (1.73 ms per video).

Test Prompt (High Motion): *"FPV drone shot flying through a futuristic sci-fi tunnel at high speed. Repeating neon ring lights, metallic walls, motion blur, symmetrical composition, cinematic lighting, 4k."*

(a) Untuned (Static Mode): γ=3.0, λ=0.1, θ_m=0.75, $\mathbf{r}$=[$\mathbf{0.2}$, $\mathbf{0.1}$]



(b) BO-Optimized (Dynamic Mode): γ=1.6, λ=0.9, θ_m=0.55, $\boldsymbol{\tau}$=[-2.0, 2.4]

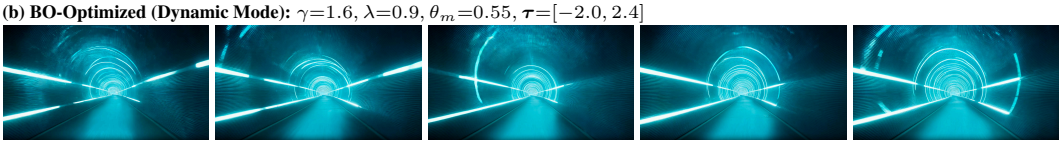


Figure 4: **Visual Ablation of Parameter Tuning and Mode Selection. (a)** The *Static-ratio* mode with naive heuristic parameters ($r_{\text{near}} = 0.2$, $\gamma = 3.0$, which widens the radial window) imposes a rigid sparsity pattern. In this high-speed FPV scene, its low retention of candidate links fails to preserve enough useful temporal connections, causing the **repeating neon rings to flicker and the tunnel geometry to distort** across frames. **(b)** The *Dynamic-threshold* mode configured by our pipeline. The **Semantic Router** classifies this as a high-motion regime, triggering the retrieval of BO-optimized parameters (relaxed $\tau_{\text{near}} = -2.0$, denser $\lambda = 0.9$) that better preserve the **symmetrical structure and smooth motion blur** in this example while balancing reconstruction fidelity and sparsity. (Notation: $\mathbf{r} = [r_{\text{near}}, r_{\text{far}}]$ denotes retention ratios; $\boldsymbol{\tau} = [\tau_{\text{near}}, \tau_{\text{far}}]$ denotes attention score thresholds).

Table 5: Measured deployed Dynamic-control overhead per attention invocation on H100/Wan-72f. The total includes QK scoring; no projected kernel-fusion benefit is counted.

Component	Latency (ms)
Dynamic QK proxy scoring	0.39
Thresholding and token-to-block aggregation	0.74
CSR/BSR index construction	0.31
FlashInfer sparse planning	0.27
Mask construction, indexing, and planning	1.32
Total Dynamic-control overhead	**1.71**

Complete-generation timing independently uses synchronized host wall-clock measurements on five fixed prompts, three runs per prompt, with matched prompt-wise seeds, checkpoint, precision, resolution, 50 sampling steps, and CFG 5.0. Dense takes 1680 ± 15s and DynamicRad 980 ± 9s ($1.71\times$). This includes text encoding, routing, sparse control and execution, all denoising steps, host orchestration, and VAE decoding. On A800, the corresponding Wan-72f pipeline takes 2480 ± 30s dense and 1530 ± 20s Dynamic ($1.62\times$). H100/FlashInfer is the primary controlled comparison; A800 establishes transfer within the tested NVIDIA CUDA stack.

E Additional Long-Context Baselines

Table 1 includes the long-context PA and SVG comparisons requested in review. SVG is our reproduction of its dense-warm-up profiling and per-head spatial/temporal pattern assignment, using the same adapted checkpoint, prompts, seeds, resolution, 50 steps, CFG, precision, hardware, and block-sparse backend as the matched long-context runs. Its latency includes warm-up, profiling, assignment, mask construction, and sparse execution. PA uses its deterministic power-of-two connectivity under the same protocol and a sparsity budget matched to Radial Attention. These results show that both methods can run at long context.

We also reproduced AdaSpa under the same H100 timing protocol and block-sparse backend, with block size 64, approximately 80% sparsity, preserved text-related sink interactions, row-wise balanced

selection, head-adaptive hierarchical allocation, ten dense warm-up steps, search at step 10, and an LSE-cached update at step 30. Timing includes warm-up, online search, LSE computation, updates, mask construction, and sparse attention. The 241f result extends the published procedure and is not an official AdaSpa result.

Table 6: AdaSpa (our reproduction) on HunyuanVideo under the matched protocol.

Frames	VR	S.C.	A.Q.	I.Q.	Latency (s)
120	0.141	0.963	0.648	0.675	958
241	0.126	0.963	0.643	0.668	420

AdaSpa reaches 80.0%/80.1% sparsity and 1.78×/1.96× speedup at 120f/241f. At 241f, DynamicRad Dynamic is 42s faster with higher quality metrics; Static is the lower-latency endpoint.

F Proxy and Router Validation

Proxy MSE versus video quality. We evaluate 48 Dynamic-threshold configurations, 24 each for low and high motion, at 78%–82% final executed block sparsity. Each uses three fixed prompts from its regime with matched prompt-wise seeds, checkpoint, CFG, sampling steps, precision, and resolution. Sparsity is computed from the final BSR index after token-to-block aggregation, dense intra-frame insertion, and fallback. Table 7 correlates mean attention-reconstruction MSE with video metrics across configurations.

Table 7: Correlation between proxy MSE and video quality at matched sparsity.

Regime	Metric	Pearson r	Spearman ρ
Low	VisionReward	-0.49	-0.46
Low	A.Q.	-0.52	-0.49
Low	I.Q.	-0.58	-0.54
Low	M.S.	-0.45	-0.42
High	VisionReward	-0.62	-0.58
High	A.Q.	-0.64	-0.60
High	I.Q.	-0.71	-0.66
High	M.S.	-0.68	-0.63

Within this range, lower proxy MSE is associated with better scores; final sparsity has absolute Spearman correlation at most 0.19 with these metrics. Across all 48 Dynamic-threshold configurations, partial Spearman correlations controlling for executed sparsity are -0.39 to -0.51 for low motion and -0.54 to -0.64 for high motion across VisionReward, A.Q., I.Q., and M.S. We use proxy MSE as a ranking signal in this tested setting, not as a universal quality predictor.

Radial support and routing. At the same 83.6% final block sparsity on Hunyuan-241f, Table 8 separates the radial prior from the selector. Global top-k scores the full grid; DynamicRad scores only radial candidates.

Table 8: Mask ablation at matched 83.6% sparsity on Hunyuan-241f.

Mask	Latency (s)	VR	S.C.
Random mask	331	0.108	0.935
Local temporal window	334	0.116	0.947
Global top-k (no radial)	472	0.129	0.966
DynamicRad Dynamic	378	0.131	0.969

The frozen router is evaluated on 2,000 held-out OpenVid-1M video–text pairs with RAFT-derived motion labels. Its row-normalized confusion matrix is shown in Table 9; balanced accuracy is 88.8%, and most errors are between neighboring regimes.

Table 9: Router confusion matrix (true regime by row, predicted regime by column).

	Low	Mid	High
Low	91.2%	8.1%	0.7%
Mid	8.4%	84.8%	6.8%
High	1.0%	8.6%	90.4%

On Hunyuan-241f, Random/Predicted/Oracle routing gives VR 0.120/0.127/0.128 and S.C. 0.952/0.965/0.967, at comparable 84.8%/84.7%/84.9% sparsity and 346/342/350s latency. Random uniformly selects among the same three BO policies over 100 assignments; Oracle uses the RAFT-derived label. Predicted routing approaches the non-deployable oracle and exceeds random assignment at similar cost.

Using a fixed pool of 43 candidate configurations in a separate proxy-drift experiment at ten drift levels, three and five contiguous motion buckets each select three unique configurations and yield mean normalized MSE 0.45626. Ten buckets select four configurations and yield 0.45613, a 0.029% decrease. We retain three regimes: finer bucketing contributes little in this tested proxy setting. The BO configurations jointly couple mode, decay, long-range, threshold, and retention parameters; arbitrary interpolation of these vectors need not preserve the calibrated sparsity budget.

G Additional Video-Quality Evaluation

Table 10: VBench Total and Dynamic Degree (DD) at long context. High-motion DD uses a RAFT-defined subset. Full-FT Dense is the checkpoint- and protocol-matched dense reference.

Setting	Method	Total (%)	DD (%)	High-motion DD (%)
Hunyuan-241f	Full-FT Dense	80.10	66.2	78.2
	Radial	80.03	65.8	77.6
	Static	79.91	64.9	76.6
	Dynamic	80.24	67.0	79.0
	Routed	80.18	66.7	78.6
Wan-145f	Full-FT Dense	79.78	64.8	76.4
	Radial	79.70	64.4	75.9
	Static	79.55	63.7	75.0
	Dynamic	79.95	65.6	77.3
	Routed	79.88	65.2	76.9

Static incurs a small dynamicity trade-off, while Dynamic and Routed preserve matched-dense-level motion. Under our local pipeline, the native checkpoints score 81.82 Total at Hunyuan-120f and 82.26 at Wan-72f, versus 80.10/79.78 for Full-FT Dense at 241f/145f. The 1.72/2.48-point differences reflect the longer horizon and finite target-length adaptation, and public native-length scores use different evaluation settings. All matched long-context methods share checkpoint, prompts and seeds, frame count, generation settings, precision, and metric pipeline.

H Additional Qualitative Results

In this section, we provide a comprehensive gallery of generated samples to demonstrate the robustness of DynamicRad across two state-of-the-art video diffusion backbones: Wan2.1-14B and HunyuanVideo.

For a controlled visual comparison, we compare a dense reference against DynamicRad. The gallery label "Original (Dense)" denotes the native dense checkpoint at standard length and checkpoint-matched Full-FT Dense at long length, not the Original-LX extrapolation row of Table 1. Both methods use identical random seeds, prompts, sampling steps, CFG scale, and backbone checkpoints within each matched setting.

Unless otherwise specified, DynamicRad results in this appendix use the **Dynamic mode with Mask-Aware LoRA (MA-LoRA)**, which is the sparsity-aware refinement introduced in Sec. 3.3. The gallery illustrates visual fidelity and temporal coherence; efficiency results for the backbone-frozen policy and the separate MA-LoRA ablation are reported in Tables 1 and 3, respectively.

Generation Configuration. To validate performance stability across varying temporal horizons, we generate samples at **High Definition (**1280×720**)** resolution using **50 sampling steps** and a CFG scale of 5.0. Specifically, we evaluate both models under standard and long-context settings:

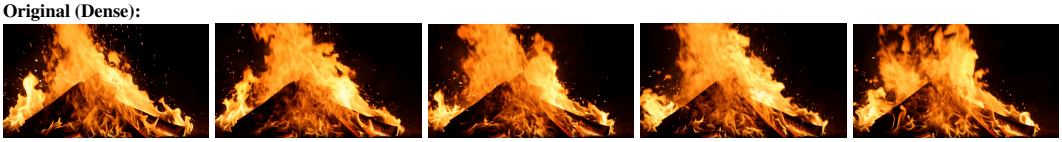
- **HunyuanVideo:** Evaluated at both **120 frames** (Standard) and **241 frames** (Long-Context, $>$200k tokens).
- **Wan2.1-14B:** Evaluated at both **72 frames** (Standard) and **145 frames** (Long-Context).

Each strip below contains **5 frames** sampled uniformly from the generated video to visualize the temporal progression.

H.1 Results on Wan2.1-14B

Case 1 (Fluid Dynamics): *"Close-up of a raging campfire at night, sparks flying upwards, dynamic flames flickering, realistic texture, cinematic lighting, 4k, high frame rate."*

Original (Dense):

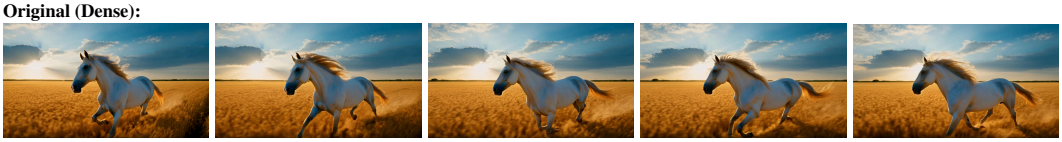


DynamicRad (Ours):



Case 2 (Fast Motion): *"A majestic white horse galloping through a field of golden wheat, wind blowing through its mane, dynamic camera angle tracking the horse, sunlight breaking through clouds, 4k, slow motion."*

Original (Dense):



DynamicRad (Ours):



Case 3 (Stylized/Anime): *"Anime style, Makoto Shinkai style, a young girl standing on a train station platform, cherry blossoms falling, vibrant blue sky with fluffy clouds, high quality animation."*

Original (Dense):

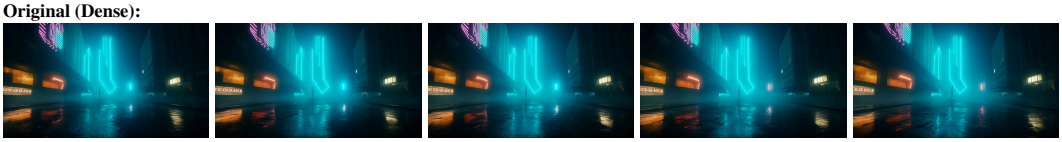


DynamicRad (Ours):

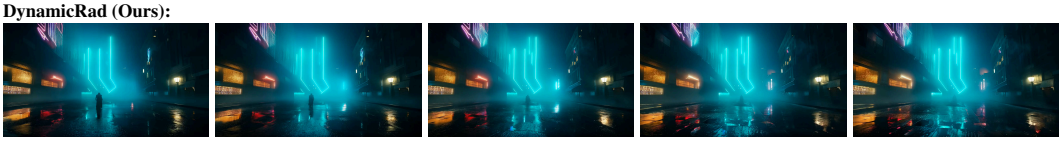


Case 4 (Complex Scene): *"Cyberpunk city street at night, neon lights reflecting on wet pavement, a mysterious figure walking through steam and rain, cinematic lighting, volumetric fog, highly detailed, 8k."*

Original (Dense):



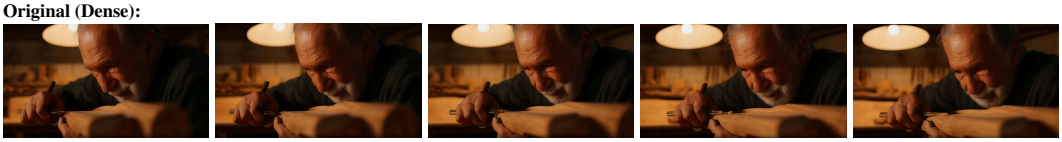
DynamicRad (Ours):



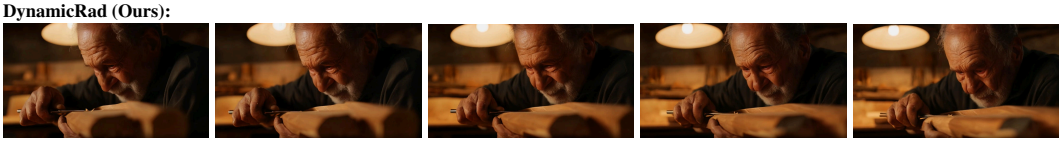
H.2 Results on HunyuanVideo

Case 5 (Portrait/Low Motion): *"A cinematic close-up of an elderly craftsman carving wood, detailed wrinkles, focused expression, warm workshop lighting, 4k, slow camera movement."*

Original (Dense):

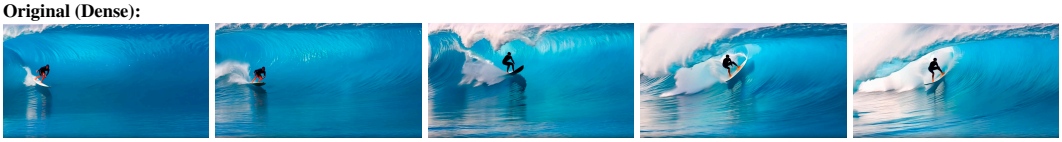


DynamicRad (Ours):

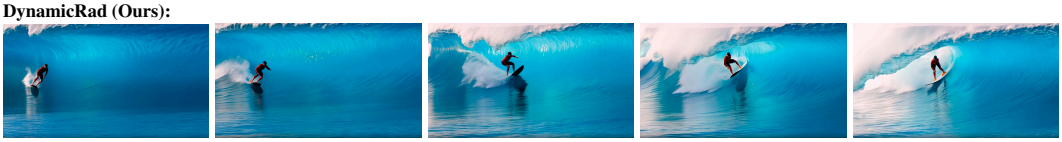


Case 6 (High Motion/Water): *"A professional surfer riding a massive blue wave, water spray, dynamic camera tracking, high speed, sunlight reflection, 4k."*

Original (Dense):

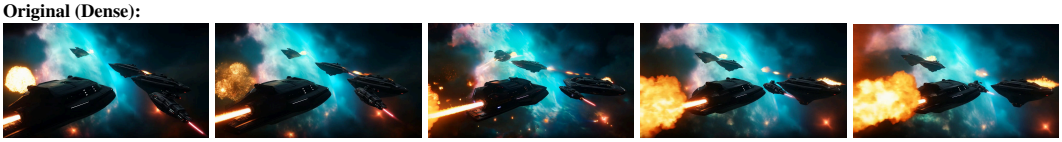


DynamicRad (Ours):

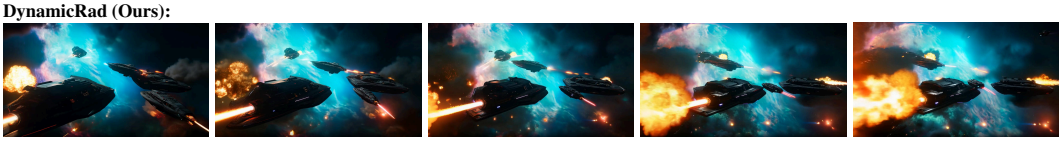


Case 7 (Sci-Fi/Texture): *"A spaceship fleet engaging in a battle near a nebula, lasers firing, explosions, colorful gas clouds, high detail, sci-fi movie style."*

Original (Dense):



DynamicRad (Ours):



Case 8 (Urban/Timelapse): *"Time-lapse of a busy City intersection at night, car trails, crowds moving fast, neon signs flashing, 4k, high quality."*

Original (Dense):



DynamicRad (Ours):



I Broader Impacts and Existing Assets

Broader impacts. DynamicRad improves the inference efficiency of long-video diffusion models by reducing the attention-side computation and memory overhead. This can have positive impacts by lowering the computational cost of long-video generation research, enabling broader experimentation with resource-constrained hardware, and reducing the energy cost associated with repeated inference and ablation studies. At the same time, more efficient long-video generation may lower the barrier to producing synthetic videos, which could be misused for misleading media, impersonation, copyright-violating content, or disinformation. DynamicRad does not introduce a new generative backbone or release a new high-risk pretrained video generator; however, it can accelerate existing video generation systems. We therefore recommend that deployments using this technique follow the safeguards of the underlying generative model, including provenance tracking, watermarking where applicable, access control for high-risk use cases, and content-safety policies for generated media.

Existing assets. We use existing models, datasets, evaluation tools, and software packages according to their respective licenses and terms of use. The licenses and usage terms were checked from official repositories or model/dataset cards at the time of submission. Table 11 summarizes the main external assets used in this work. We do not redistribute the original model weights, datasets, or benchmark assets as part of this submission; users should obtain them from the corresponding official sources and comply with their licenses and usage terms.

Table 11: Existing assets used in this work.

Asset	Type	Usage	License / Terms
HunyuanVideo	Model/backbone	Long-video diffusion backbone for evaluation	Tencent Hunyuan Community License Agreement; includes territorial restrictions; third-party components follow their original licenses
Wan2.1-14B	Model/backbone	Video diffusion backbone for evaluation	Apache-2.0
OpenVid-1M	Dataset	Router supervision and Mask-Aware LoRA fine-tuning subset	CC-BY-4.0 for dataset metadata; dataset card states intended research/non-commercial use and requires compliance with original source licenses
VBench	Evaluation benchmark	Subject consistency, aesthetic quality, and imaging quality evaluation	Apache-2.0
VisionReward	Evaluation model/code	Human-aligned video quality scoring	Code: Apache-2.0; model weights follow the Hugging Face model card / upstream CogVLM2-based checkpoint terms
FlashInfer	Software/kernel library	Block-sparse attention kernel support and inference acceleration	Apache-2.0
RAFT	Optical-flow model/code	Optical-flow pseudo-labeling for semantic motion router supervision	BSD-3-Clause
PEFT / LoRA	Software library	Parameter-efficient Mask-Aware LoRA adaptation	Apache-2.0

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main methodological contributions and the intended scope of the paper. These claims are supported by the experimental results in Secs. 4.1–4.3 and the stated limitations in Sec. 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations explicitly in Sec. 5 under "Limitations and Future Work." These include extreme fast-motion cases, ambiguous prompts, and hardware generalization issues.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not claim a formal complexity bound for the implemented method. Appendix A gives an explicitly idealized scaling illustration for the radial-decay prior, with its assumptions and short derivation stated there; no theorem/proof contribution is presented.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides the method description, experimental setup, model backbones, hardware setup, BO search description, and evaluation metrics in the main paper and appendix. A code link is also provided to support reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code and usage instructions for the Wan implementation are publicly available. The current release does not include all materials needed to reproduce every main experimental result.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setting is described in Sec. 4.1, including the backbones, frame settings, hardware environment, BO-based configuration pipeline, and LoRA setup. Additional hyperparameter ranges and BO details are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The five-prompt end-to-end latency study reports standard deviations across prompt-level means after three timing runs per prompt. The video-quality comparisons do not include formal significance tests; conclusions are supported by matched protocols, two backbones, and ablations.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sec. 4.1 reports the main compute environment, including single-H100 and 8×H100 settings, and the paper reports representative runtime and profiling information. Additional BO efficiency and routing overhead details are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our knowledge, the research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss both positive and negative societal impacts in Appendix I. The positive impacts include reduced computational cost and broader accessibility for long-video generation research, while the negative impacts include potential misuse for synthetic media or disinformation.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new pretrained generative model weights or new scraped datasets. The contribution provides an implementation of the method rather than a new high-risk generative model asset.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The external models, datasets, benchmarks, and software packages used in this work are cited in the main paper, and their licenses or usage terms are summarized in Appendix I, Table 11. We do not redistribute the original external assets and use them according to their respective licenses and terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper provides a new implementation of the proposed method together with documentation and usage instructions in the accompanying code repository and supplemental material.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper reports no crowdsourcing or research involving human participants.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper reports no research involving human participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not an important, original, or non-standard component of the proposed method. The core contribution is a sparse-attention and routing mechanism for long video diffusion.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.