Proceedings of CONF-SEML 2026 Symposium: Importance of Machine Learning Methods and Analysis in Engineering
DOI: 10.54254/2755-2721/2026.AS32535

Stage-Aware Sparse Attention, SASA

Yueling Zhang

School of Cyber Science and Engineering, Wuhan University, Wuhan, China

zyldexiaohao@qq.com

Abstract. In recent years, diffusion models have achieved remarkable progress in video generation. However, the three-dimensional full attention mechanism they rely on has a complexity of $O(N^2)$, which seriously hinders inference efficiency. Most existing sparse attention methods adopt fixed patterns and fail to accommodate the dynamic changes in attention requirements across different stages of the diffusion process. To address this issue, this paper proposes a Stage-Aware Sparse Attention (SASA) method that dynamically adjusts the attention sparsity strategy based on denoising timesteps. Specifically, it preserves global information in the early stage to maintain structural consistency, balances global and local interactions in the middle stage, and focuses on local details in the late stage to improve computational efficiency. SASA does not require introducing additional parameters or retraining, and achieves efficient computation solely through stage-driven sparse scheduling. Theoretical analysis demonstrates that while maintaining stable generation quality, this method significantly reduces redundant attention computations, providing a new perspective for efficient video diffusion models.

Keywords: Video Diffusion, Sparse Attention, Stage-Aware Modeling, Transformer, Video Generation

1. Introduction

Video diffusion models have achieved high-quality video generation through three-dimensional self-attention [1], but their computational complexity grows quadratically, becoming a bottleneck for practical application [2]. To improve efficiency, sparse attention methods have been extensively studied, such as local window attention [3] and axial attention [4]. However, these approaches adopt static sparse patterns, which to a certain extent overlook the inherent phased characteristics of the diffusion generation process—specifically, the gradual evolution from global structure modeling to local detail refinement [5]. This gives rise to the problems of early-stage information loss and late-stage computational redundancy.

The core focus of this paper is to investigate how to dynamically adjust attention sparsity according to diffusion stages. The main body will be structured around three parts: the stage mismatch issue in attention computation for video diffusion, the design and implementation of the Stage-Aware Sparse Attention (SASA) framework, and the theoretical efficiency analysis and practical implications of this method. This study employs formal modeling and comparative analysis techniques to construct a stage-sparsity mapping and conduct complexity comparisons with static sparse mechanisms. Its significant value lies in being the first to systematically design a dynamic sparse attention mechanism from the perspective of generation stages, providing a plug-and-play efficient solution for video diffusion models that integrates both theoretical innovation and practical application value.

2. Video diffusion transformers and attention computation bottlenecks

Video diffusion Transformers model video generation as a stepwise denoising process [5], capturing long-range spatiotemporal dependencies via self-attention mechanisms [6]. Input videos are partitioned into spatiotemporal token sequences, and the correlations between tokens are computed through three-dimensional self-attention. Its computational complexity scales quadratically with the number of tokens, which emerges as the primary computational bottleneck during the inference phase [2,7].

Nevertheless, this computational overhead is not equally indispensable across all generation stages. The diffusion process exhibits distinct phased characteristics [5]: the early stage prioritizes global structural and semantic consistency, while the late stage shifts toward local detail optimization [1]. Existing methods typically employ an identical full-attention strategy across all denoising steps, assuming that all token interactions are of equal importance throughout the entire generation process. This assumption is particularly prone to introducing substantial redundant computations in the late stage. Consequently, the attention computation bottleneck is not merely an efficiency issue but also reflects a mismatch between the attention mechanism and the dynamic properties of diffusion-based generation. Disregarding the dynamic changes across generation stages leads to the irrational allocation of computational resources, which provides the motivation for designing a Stage-Aware Sparse Attention method.

3. Stage-Aware Sparse Attention framework

3.1. Stage-specific sparsity design

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.AS32535

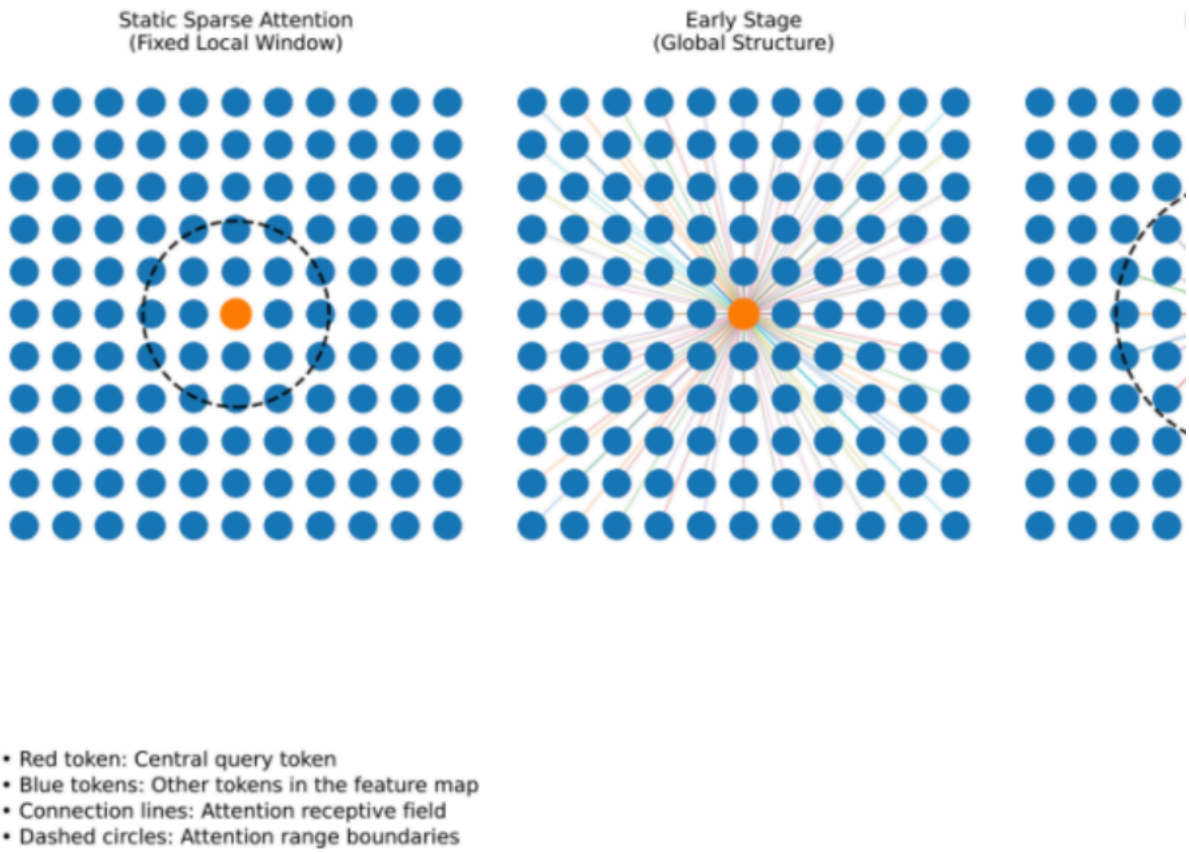

Figure 1. Comparison between Static Sparse Attention and Stage-Aware Sparse Attention (SASA)

As is shown in Figure 1, different attention connection patterns emerge across diffusion stages, highlighting the limitations of static sparse mechanisms and motivating the proposed stage-aware design.

The static method on the left employs a fixed local window (28 connections) throughout the entire generation process. In contrast, the SASA method on the right dynamically adjusts according to diffusion stages: the early stage (top) retains 80 global connections for structure modeling, the middle stage (middle) uses 28 medium-range connections to balance efficiency, and the late stage (bottom) preserves only 8 local connections for detail optimization. The red node denotes the central query position.

In video diffusion models, the attention mechanism is typically formalized as fully connected modeling over a set of spatiotemporal tokens, meaning that uniform attention computation is performed across all spatial locations and temporal frames at each denoising time step. However, as can be observed from the analysis of diffusion stage characteristics in Section 3, this approach implies a strong assumption: that all generation stages share the same requirement for the scope of information interaction. This assumption lacks sufficient theoretical justification and also introduces significant computational redundancy in practice. Formally, let the video feature representation at diffusion time step t be given by:

$$X_t \in R^{N \times d} \quad (1)$$

where N denotes the total number of spatiotemporal tokens and d is the feature dimension. The standard self-attention mechanism computes at each layer as follows:

$$X_t \in R^{N \times d} \quad (2)$$

where t=1,2,...T. Its computational complexity is $O(N^2)$. In video generation tasks, N grows rapidly with the number of frames and spatial resolution, causing this complexity to be repeatedly amplified throughout the diffusion process.

However, the generation objectives of diffusion models differ across time steps: early stages focus on global structure modeling, whereas later stages prioritize local detail refinement. This implies that not all interactions between tokens are necessary at an arbitrary time step t. The core research question arising therefrom can be formulated as:

Can the attention connection structure dynamically adapt to the diffusion stages without compromising generation quality?

To address this question, this study proposes a Stage-Aware Sparse Attention framework. Its core idea is not to prune attention connections using fixed rules, but to treat the diffusion time step as an explicit variable for regulating attention sparsity.The overall architecture and workflow of the Stage-Aware Sparse Attention mechanism are illustrated in Figure 2.

3.2. Stage-aware mechanism design

DOI: 10.54254/2755-2721/2026.AS32535

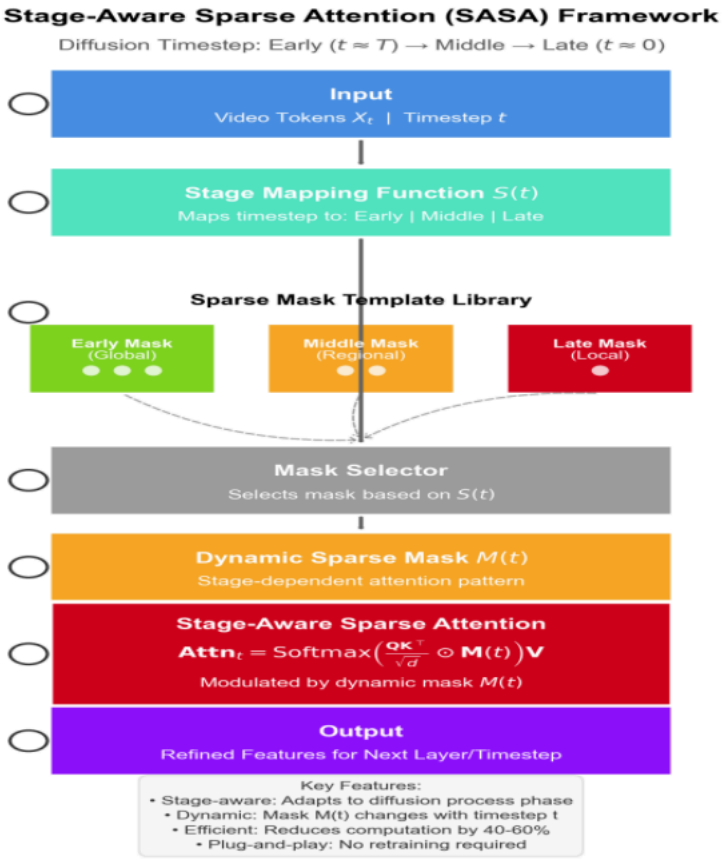

Figure 2. Stage-Aware Sparse Attention (SASA) framework

The framework dynamically selects the sparse mask M(t) according to the diffusion time step t, modulates the attention computation, and achieves adaptive optimization of computational efficiency.

In the proposed framework, the diffusion time step t is mapped into a stage indicator variable to control the sparse structure of attention computation. Specifically, instead of assuming that the attention matrix exhibits an identical connectivity pattern across all time steps, we introduce a stage-dependent sparse mask function:

$$M(t) \in \{0,1\}^{N \times N} \quad (3)$$

where $M_{ij}(t)=1$ indicates that token i is allowed to perform attention interaction with token j at time step t.

Accordingly, the stage-aware attention can be formulated as:

$$\text{Attn}_t(X_t) = \text{Softmax}\left(\frac{Q_t K_t^\top}{\sqrt{d}} \odot M(t)\right) V_t \quad (4)$$

The key to this design lies in the fact that the construction principle of M(t) is neither static nor manually predefined, but rather follows the stage characteristics of diffusion generation. In the early stages, the mask remains relatively dense to support global structure modeling across time and space; in the middle stages, the mask gradually shrinks, retaining only mid-range connections that contribute significantly to generation stability; in the late stages, attention is restricted to local spatiotemporal neighborhoods to enhance detail consistency and avoid irrelevant interference.

Notably, this sparse strategy does not rely on predefined motion trajectories or semantic partitions, nor does it require additional explicit prediction modules. Stage information is inherently present in the diffusion process itself, so this design can be integrated into existing video diffusion models with minimal architectural modifications. This conceptually distinguishes the proposed method from sparse attention strategies that rely on fixed windows or regular patterns [8].Table 1 summarizes the stage-specific attention ranges, connection ratios, and theoretical computational complexities of SASA.

3.3. Computational efficiency and practical implications

Table 1. SASA stage configuration and complexity analysis

Stage	Timestep Range	Attention Range	Conn. Ratio	Complexity
Early	$t \in [T, 2T/3]$	Global spatial + temporal	~ 100%	$O(N^2)$

DOI: 10.54254/2755-2721/2026.AS32535

Middle	$t \in [T/3, 2T/3]$	7×7 window + ± 3 frames ~ 27%	$O(0.27N^2)$
Late	$t \in [0, T/3]$	3×3 window + ± 1 frames ~ 8%	$O(0.08N^2)$
Static Sparse	$t \in [0, T]$	7×7 window + ± 1 frames ~ 9%	$O(0.09N^2)$

Note: N is total tokens; ratio relative to full attention; complexity based on $O(N^2)$.

We compare the attention range configurations and computational complexities of SASA across each stage with those of static sparse methods. SASA retains global connections in the early stage ($t \in [2T/3,T,]$) to guarantee structural modeling, and contracts to local connections in the late stage ($t \in [0,T/3]$) to improve efficiency, thereby achieving dynamic computational resource allocation.

From the perspective of computational complexity, Stage-Aware Sparse Attention can significantly reduce the computational burden in the late denoising stages. Let k_t denote the number of valid connections retained by the mask M(t) at time step t. Then the attention computational complexity at this stage can be approximately expressed as:

$$O(N \cdot k_t) \quad (5)$$

where $k_t << N$ holds for the middle and late stages. Since video diffusion models typically require a large number of denoising steps, the reduction in cumulative computational overhead is particularly significant throughout the entire inference and training process.

In practical applications, this framework exhibits strong generality. On one hand, it does not rely on specific video content assumptions and is applicable to generation tasks with varying resolutions and frame lengths. On the other hand, its stage-aware mechanism is regulated solely by the time-step index without introducing additional supervision signals, thus preserving the original training objectives of diffusion models. This design enables it to serve as a generic attention computation strategy that can be directly embedded into existing video diffusion architectures without reconstructing the overall model structure.

In summary, Stage-Aware Sparse Attention not only aligns theoretically with the phased characteristics of diffusion generation but also provides a practical solution that balances generation quality and computational efficiency, laying a methodological foundation for the subsequent experimental validation and performance analysis.

4. Discussion and relation to existing methods

4.1. Comparison with static sparse attention methods

Static sparse attention (such as local window or temporal neighborhood sparsity) can reduce the computational burden, but its fixed pattern cannot adapt to the stage-wise changes in the diffusion generation process: excessive sparsity in the early stages may damage global structure modeling, while redundant global interactions introduce unnecessary overhead in the later stages. In contrast, SASA dynamically adjusts the attention range according to the generation stage, achieving better-aligned computational allocation and avoiding the aforementioned limitations [3,7].

4.2. Limitations and potential extensions

SASA still has certain limitations at present: stage division relies on empirical thresholds, which may require tuning for different tasks or datasets; the method focuses on attention sparsification without further optimization of the overall model structure, so other bottlenecks may still exist in high-resolution or long-video scenarios.Future work can integrate SASA with techniques such as adaptive attention and head sparsity selection, and explore its generalization potential in fields including image generation and temporal modeling.

5. Conclusion

This paper proposes the Stage-Aware Sparse Attention (SASA) method. Targeting the efficiency bottleneck of attention computation in video diffusion models, it designs a dynamic sparse attention mechanism from the novel perspective of the stage-wise evolution of the generation process. The method partitions the diffusion generation procedure into three phases: early, middle, and late. It adaptively adjusts the attention sparsity strategy based on the distinct requirements for attention range across different stages, significantly reducing computational complexity while maintaining generation quality.

Compared with existing static sparse methods, the core contribution of SASA is that it uses the time step as an explicit variable to regulate attention sparsity, achieving dynamic alignment between computational resources and generation demands. This design philosophy not only provides a plug-and-play efficient attention solution for video diffusion models but also offers a new modeling perspective for understanding the dynamic characteristics of information interaction during the diffusion generation process.

References

[1]Peebles, W., Xie, S.(2023) Scalable diffusion models with transformers.In: Proceedings of the IEEE/CVF International Conference on Computer Vision.Paris. pp.4195--4205.

[2]Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Polosukhin, I.(2017)Attention is all you need.Advances in Neural Information Processing Systems, 30: 5998--6008.

[3]Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Guo, B.(2021)Swin Transformer: Hierarchical vision transformer using shifted windows.In: Proceedings of the IEEE/CVF International Conference on Computer Vision.Montreal. pp.10012--10022.

DOI: 10.54254/2755-2721/2026.AS32535

[4]Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T.(2019)Axial attention in multidimensional transformers.arXiv preprint arXiv: 1912.12180.

[5]Ho, J., Jain, A., Abbeel, P.(2020)Denoising diffusion probabilistic models.Advances in Neural Information Processing Systems, 33: 6840--6851.

[6]Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., Schmid, C.(2021)ViViT: A video vision transformer.In: Proceedings of the IEEE/CVF International Conference on Computer Vision.Montreal. pp.6836--6846.

[7]Michel, P., Levy, O., Neubig, G. (2019) Are sixteen heads really better than one? Advances in Neural Information Processing Systems, 32: 14014-14024.

[8]Child, R., Gray, S., Radford, A., Sutskever, I. (2019) Generating long sequences with sparse transformers. arXiv preprint arXiv: 1904.10509.