Reducing Attention Distribution Error with Unified Tail Aggregation for Sparse Attention

Hyunwoo Yu [*1] **Jongbeom Lee** [*1] **Jaemin Choi** [*1] **Jincheol Yang** [*1] **Yubin Cho** [*1] **Joonsik Nam** [1]
Seunghun Moon [1] **Jung-Woo Chang** [2] **Bongjoon Hyun** [3] **Kyomin Sohn** [3] **Kyeongbo Kong** [†4] **Suk-Ju Kang** [†1]

Abstract

While Diffusion Transformers (DiTs) deliver remarkable visual quality in video generation, the massive computational overhead of 3D spatio-temporal attention limits their scalability. To evaluate and optimize sparse attention mechanisms, existing studies predominantly rely on the Top-K Oracle Policy. However, this approach employs rigid truncation that naively discards the continuous tail of the attention distribution, introducing structural errors that degrade temporal consistency during iterative diffusion processes. To address this fundamental flaw, we provide an oracle analysis of these distributional shifts and introduce a novel *Low-Delta* Oracle Policy. Building on a mathematical proof demonstrating that sparse attention achieves zero error when grouping identical attention scores, our approach prioritizes the structural integrity of the entire attention distribution. As a promising correction mechanism, we propose a Unified Tail Aggregation (UTA) method. By aggregating logits where the score variance is bounded by a marginal delta, UTA supplements a single aggregated logit to restore the attention distribution. Extensive empirical evaluations demonstrate that our approach significantly outperforms the Top-K oracle, achieving up to a 97.4% reduction in mean squared error (MSE) at a 50% sparsity level. By establishing a tighter theoretical upper bound, this work provides a rigorous foundation for evaluating and stabilizing future sparse attention systems.

[1]Department of Electronic Engineering, Sogang University, Seoul, Republic of Korea [2]University of California San Diego, La Jolla, CA, USA [3]Samsung Electronics, Hwaseong-si, Republic of Korea [4]Department of Electrical & Electronics Engineering, Pusan National University, Republic of Korea. Correspondence to: Kyeongbo Kong <kbkong@pusan.ac.kr>, Suk-Ju Kang <sjkang@sogang.ac.kr>.

International Conference on Machine Learning, AdaptFM: Resource-Adaptive Foundation Model Inference, Seoul, South Korea. *Copyright 2026 by the author(s).*

1. Introduction

While Diffusion Transformers (Peebles & Xie, 2023) have achieved remarkable success in image synthesis, extending their capabilities to video generation (Wan et al., 2025; Kong et al., 2024; Yang et al., 2024; Brooks et al., 2024; HaCohen et al., 2024) introduces massive computational overhead. Although adapting 3D spatio-temporal full attention produces visually compelling and consistent outputs, the exponential increase in token counts across the temporal dimension renders training and inference computationally prohibitive. To mitigate this burden, various sparse attention mechanisms aim to determine which key and value tokens are most critical for a given query. While static methods (Xi et al., 2025b; Zhang et al., 2025c; Li et al., 2025b) reduce complexity through predetermined attention masks, dynamic methods (Zhang et al., 2025b; Xu et al., 2025; Xia et al., 2025) preserve relevant information by evaluating sparsity patterns during inference.

To evaluate these strategies and establish a theoretical performance upper bound, existing studies predominantly rely on Top-K or Top-P oracle policies. However, we identify a limitation in this approach: as sparsity increases, Top-K relies on rigid truncation that discards the continuous tail of the attention distribution, distorting the inherent probability distribution of full attention. In iterative diffusion processes, this distribution shift acts as a compounding error across multiple timesteps, ultimately degrading the temporal consistency and visual quality of generated videos.

We argue that a true upper bound for sparse attention must shift away from naive truncation, and instead prioritize preserving the structural integrity of the entire attention distribution. To realize this paradigm, we first establish a mathematical proof demonstrating that, if multiple logit indices share the exact same attention score, sparse attention can be executed with zero error. Inspired by this theoretical foundation, we introduce a *Low-Delta* Oracle Policy that exploits logit indices with minimal differences in their attention scores. To empirically verify our claims, we propose Unified Tail Aggregation (UTA) as a straightforward implementation. By grouping tokens whose score variance is bounded by a marginal delta, UTA compresses redundant

contextual information while mathematically preserving the original score distribution. The main contributions are:

- We introduce a novel *Low-Delta* Oracle Policy based on a mathematical proof showing that sparse attention can achieve zero error when grouping identical attention scores, fundamentally resolving structural errors caused by rigid truncation in Top-K policies.
- We propose Unified Tail Aggregation (UTA) as a straightforward implementation of the *Low-Delta* Oracle Policy, which restores the attention distribution by supplementing a single aggregated logit.
- We demonstrate that our approach outperforms the Top-K oracle and verify that masked logits in existing sparse attention frameworks predominantly exhibit *Low-Delta* characteristics, so that applying UTA directly to masked regions improves attention recall and reduces output MSE.

2. Related Work

Sparse Attention for Video Diffusion. Sparse attention for video diffusion has been actively studied to alleviate the excessive computational overhead of attention. Static mask-based methods (Xi et al., 2025b; Zhang et al., 2025c; Li et al., 2025b;a; Chen et al., 2025a) apply predefined attention masks by analyzing attention score patterns a priori. Dynamic methods (Zhang et al., 2025b; Xu et al., 2025; Xia et al., 2025; Zhao et al., 2025; Tan et al., 2025; Liu et al., 2025; Sun et al., 2025; Zhang et al., 2025d; Wu et al., 2025; Durvasula et al., 2025; Xi et al., 2025a; Yang et al., 2025a) apply sparsity masks by analyzing the attention map's sparsity patterns during runtime. Both approaches conventionally adopt Top-K selection as their oracle policy. Recently, alternative methods determine the sparsity ratio based on probability mass via Top-P (Zhang et al., 2026; Yang et al., 2025b). However, these approaches fundamentally fail to resolve the attention distribution shift caused by omitted logits at high sparsity. While some works (Chen et al., 2025b; Zhang et al., 2025a) cache and reuse masked logits, this incurs a critical trade-off with memory or computational overhead. Departing from the conventional Top-K or Top-P paradigms, our work presents a new direction focused on identifying logits with similar attention scores and compensating for the compromised softmax distribution by aggregating masked logits into a single representative.

3. Method

3.1. Preliminaries

We first describe full attention in matrix form and then present the per-query expression. Let $Q \in \mathbb{R}^{M\times d}$, $K \in \mathbb{R}^{L\times d}$, and $V \in \mathbb{R}^{L\times d_v}$ denote the query, key, and value matrices, respectively, where M is the number of query tokens and L is the number of key/value tokens. The full attention output is

$$O \;=\; \mathrm{Attn}_{\mathrm{full}}(Q,K,V) \;\triangleq\; \mathrm{Softmax}\!\left(\frac{QK^\top}{\sqrt{d}}\right)V. \quad (1)$$

For a single query token $\boldsymbol{q}_m$, let $\mathcal{S}$ denote the index set of all tokens attending to $\boldsymbol{q}_m$. The attention logit z_i and weight $p(i)$ for $i \in \mathcal{S}$ are

$$z_i \triangleq \boldsymbol{q}_m \boldsymbol{k}_i^\top, \qquad p(i) \triangleq \frac{\exp(z_i)}{\sum_{j\in\mathcal{S}}\exp(z_j)}. \quad (2)$$

Accordingly, the per-query attention output is

$$\mathrm{Attn}_{\mathrm{full}}(\boldsymbol{q}_m) = \frac{\sum_{i\in\mathcal{S}}\exp(z_i)\boldsymbol{v}_i}{\sum_{j\in\mathcal{S}}\exp(z_j)} = \sum_{i\in\mathcal{S}} p(i)\boldsymbol{v}_i. \quad (3)$$

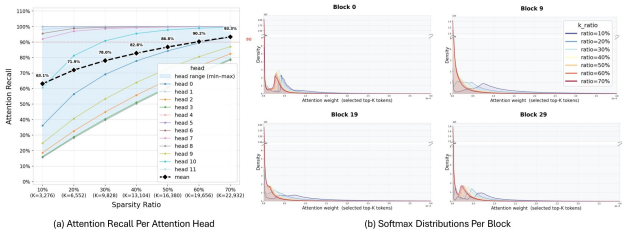


Figure 1. (a) Attention recall per attention head and (b) softmax distributions per block, as a function of the sparsity ratio. Even under an idealized oracle setting, the softmax distributions exhibit significant variation depending on the sparsity level.

3.2. The Flaw of Truncation: Distribution Collapse in Top-K

The prevailing standard for establishing a theoretical upper bound of sparse attention is the Top-K Oracle. We argue that this naive truncation suffers from a critical mathematical flaw that destroys the structural integrity of the attention mechanism, as visualized in Fig. 1.

Under rigid Top-K truncation, the full index set $\mathcal{S}$ is decomposed into the selected set $\mathcal{T}$ and the discarded set $\mathcal{D}$:

$$\mathcal{S} = \mathcal{T} \sqcup \mathcal{D}, \quad |\mathcal{T}| = K, \;\; \mathcal{D} = \mathcal{S} \setminus \mathcal{T}. \quad (4)$$

The induced weight under truncation, $p_{\mathcal{T}}(i)$, is the renormalized distribution restricted to $\mathcal{T}$:

$$p_{\mathcal{T}}(i) \triangleq \frac{p(i)}{\sum_{j\in\mathcal{T}} p(j)} = \alpha\, p(i), \;\; \alpha \triangleq \Big(\sum_{j\in\mathcal{T}} p(j)\Big)^{-1}, \quad (5)$$

for $i \in \mathcal{T}$, where $\alpha \geq 1$ inflates all surviving weights to compensate for the removed probability mass, directly inducing re-normalization distortion.

Comparing $\mathrm{Attn}_{\mathrm{full}}(\boldsymbol{q}_m)$ against the Top-K output $\mathrm{Attn}_{\mathcal{T}}(\boldsymbol{q}_m) = \sum_{i\in\mathcal{T}} p_{\mathcal{T}}(i)\boldsymbol{v}_i$, the discrepancy decomposes as:

$$\mathrm{Attn}_{\mathrm{full}} - \mathrm{Attn}_{\mathcal{T}} = \underbrace{\sum_{i\in\mathcal{T}}(1-\alpha)p(i)\boldsymbol{v}_i}_{\text{re-normalization distortion}} + \underbrace{\sum_{i\in\mathcal{D}} p(i)\boldsymbol{v}_i}_{\text{dropped-mass loss}} . \quad (6)$$

Eq. (6) shows that Top-K simultaneously distorts surviving weights via re-normalization and removes the entire tail contribution, yielding compounded errors that accumulate across iterative denoising steps in video diffusion models.

3.3. Theoretical Ideal: Zero-Error Aggregation

We first establish the theoretical foundation by analyzing the exact condition under which sparse attention incurs zero approximation error. Under the partition $\mathcal{S} = \mathcal{T} \sqcup \mathcal{D}$, the numerator and denominator of Eq. (3) separate into $\mathcal{T}$- and $\mathcal{D}$-dependent components:

$$\mathrm{Attn}_{\mathrm{full}}(\boldsymbol{q}_m) = \frac{\sum_{n\in\mathcal{T}} \exp(z_n)\boldsymbol{v}_n + \sum_{i\in\mathcal{D}} \exp(z_i)\boldsymbol{v}_i}{\sum_{n\in\mathcal{T}} \exp(z_n) + \sum_{i\in\mathcal{D}} \exp(z_i)}. \quad (7)$$

Theorem 3.1 (Zero-Error Aggregation)**.** *Let $\mathcal{S}$ be the full token set and $\mathcal{D} \subseteq \mathcal{S}$ be a truncation set, with $N_d \triangleq |\mathcal{D}|$. For a given query $\boldsymbol{q}_m$, assume that all attention logits over $\mathcal{D}$ are identical:*

$$z_i = \hat{z}_{\mathcal{D}}, \quad \forall i \in \mathcal{D}, \quad \hat{z}_{\mathcal{D}} \triangleq \frac{1}{N_d}\sum_{i\in\mathcal{D}} z_i. \quad (8)$$

Define the aggregated value $\hat{\boldsymbol{v}} \triangleq \frac{1}{N_d}\sum_{i\in\mathcal{D}} \boldsymbol{v}_i$. Then replacing $\{z_i, \boldsymbol{v}_i\}_{i\in\mathcal{D}}$ with the single representative pair $(\hat{z}_{\mathcal{D}}, \hat{\boldsymbol{v}})$, weighted by N_d, preserves the attention output exactly:

$$\mathrm{Attn}_{\mathrm{full}}(\boldsymbol{q}_m) = \frac{\sum_{n\in\mathcal{T}} \exp(z_n)\boldsymbol{v}_n + N_d \exp(\hat{z}_{\mathcal{D}})\hat{\boldsymbol{v}}}{\sum_{n\in\mathcal{T}} \exp(z_n) + N_d \exp(\hat{z}_{\mathcal{D}})}. \quad (9)$$

The proof factors $\exp(\hat{z}_{\mathcal{D}})$ out of the $\mathcal{D}$-summations in Eq. (7) and is provided in supplementary materials.

***Low-Delta* Oracle Policy.** Building on this ideal, the proposed *Low-Delta* oracle shifts the objective of sparse attention. Instead of naive Top-K truncation, it identifies and exploits indices that share similar attention scores. Grouping tokens into clusters with similar scores serves as a highly effective method to compress and utilize information while preserving the overall probability mass.

3.4. Unified Tail Aggregation (UTA)

3.4.1. Empirical Anatomy of Attention: Tail Logit Concentration.

To bridge the zero-error condition in Section 3.3 to implemental settings, we hypothesize that concentration of logits within a narrow range implies low variance and near-identical values. Consequently, approximating this concentrated region with a compact representation should substantially reduce the approximation error.

To examine this hypothesis, we analyze the attention weight distributions across various blocks and heads within a standard video diffusion model. As shown in Fig. 3(a), the distribution of the entire attention score space is highly skewed: a tiny fraction of high-attention tokens dominates the probability mass, while the vast majority of logits are concentrated near zero. In Fig. 3(b), we observe a dense cluster of small-magnitude logits when focusing on the lower 90% of tokens. Furthermore, the CDF over this lower region rises sharply (Fig. 3(c)), indicating that most tail tokens lie within a narrow score range close to zero. In addition, Fig. 3(d) shows that differences between adjacent sorted values form a sharp peak at zero, confirming that many tail logits are nearly identical.

These empirical findings provide a practical bridge to our theoretical modeling. The tail is not mere noise; it forms a dense region in which attention scores are tightly clustered around zero with negligible variance. Therefore, Top-K truncation discards a substantial portion of the distribution and risks losing collective context carried by this tail mass.

3.4.2. Unified Tail Aggregation (UTA).

Motivated by the strong tail concentration above, we present UTA in the simplest *single-bin* form, where the entire tail set $\mathcal{D}$ is aggregated into one representative token (Fig. 2). After obtaining the Top-K set $\mathcal{T}$, we aggregate $\mathcal{D}$ into a single bin with size $N_d = |\mathcal{D}|$ and statistics

$$\bar{z}_{\mathcal{D}} \triangleq \frac{1}{N_d}\sum_{i\in\mathcal{D}} z_i, \quad \hat{\boldsymbol{v}}_{\mathcal{D}} \triangleq \frac{1}{N_d}\sum_{i\in\mathcal{D}} \boldsymbol{v}_i. \quad (10)$$

Using the empirical concentration $z_i \approx \bar{z}_{\mathcal{D}}$ for $i \in \mathcal{D}$, the tail sums in Eq. (7) admit

$$\sum_{i\in\mathcal{D}} \exp(z_i) \approx N_d \exp(\bar{z}_{\mathcal{D}}) = \exp(\bar{z}_{\mathcal{D}} + \log N_d), \quad (11)$$

$$\sum_{i\in\mathcal{D}} \exp(z_i)\boldsymbol{v}_i \approx \exp(\bar{z}_{\mathcal{D}} + \log N_d)\,\hat{\boldsymbol{v}}_{\mathcal{D}}. \quad (12)$$

Following the zero-error principle, we define the compensated tail logit

$$\hat{z}_{\mathcal{D}} \triangleq \bar{z}_{\mathcal{D}} + \log N_d, \quad (13)$$

where the $\log N_d$ term naturally restores the cardinality factor inside the softmax. Concatenating this single representative with the Top-K elements yields

$$\mathrm{Attn}_{\mathrm{UTA}}(\boldsymbol{q}_m) = \frac{\sum_{n\in\mathcal{T}} \exp(z_n)\boldsymbol{v}_n + \exp(\hat{z}_{\mathcal{D}})\hat{\boldsymbol{v}}_{\mathcal{D}}}{\sum_{n\in\mathcal{T}} \exp(z_n) + \exp(\hat{z}_{\mathcal{D}})}. \quad (14)$$

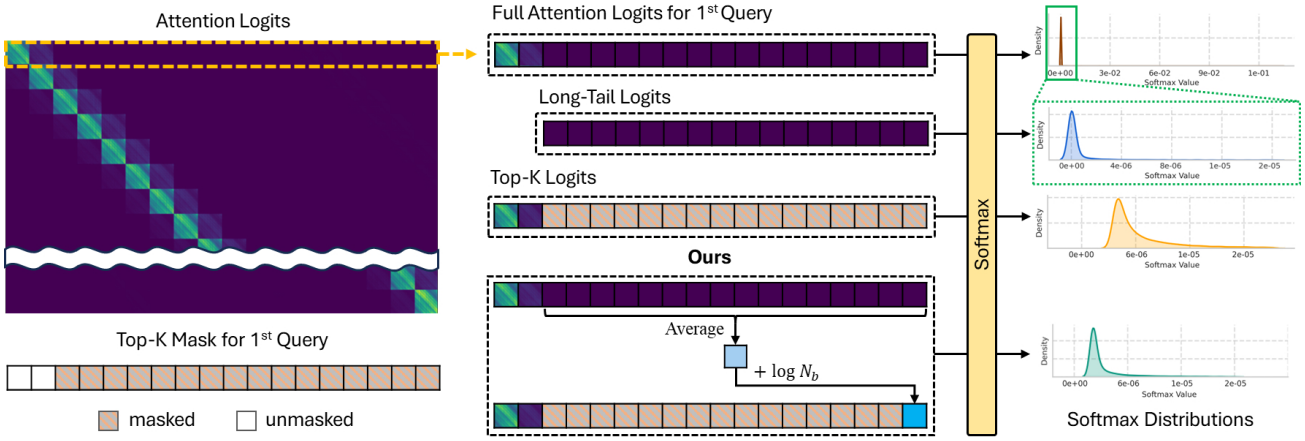


Figure 2. Overview of the Unified Tail Aggregation (UTA) framework, illustrated for the first query and simplified for clarity. Discarded tail tokens are aggregated into a single representative whose logit is compensated by $\log N_d$, restoring both the softmax denominator and the value contribution.

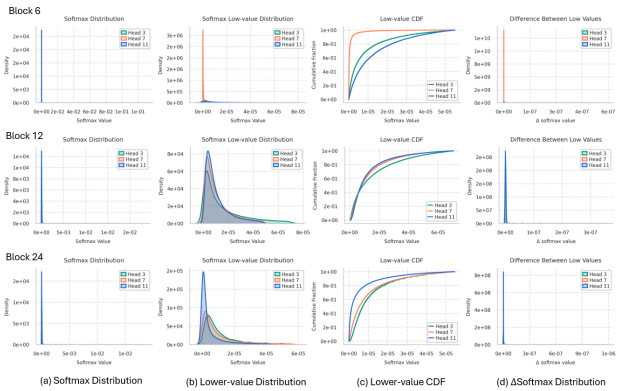


Figure 3. Analysis of attention weight distributions. (a) Softmax distributions for all tokens. (b) Softmax distributions for lower-value tokens. (c) CDF for softmax values. (d) Distributions of Δsoftmax values, showing a sharp peak at zero.

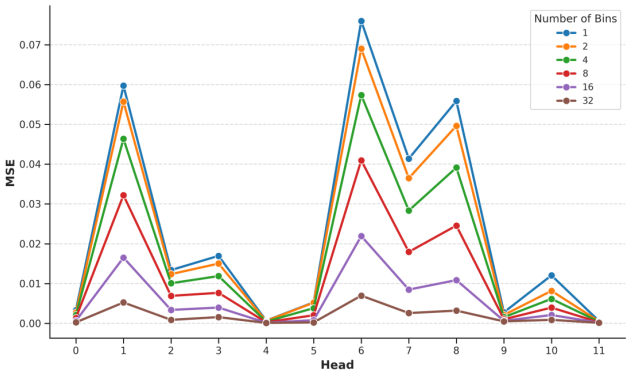


Figure 4. MSE of approximated attention outputs across heads as the number of bins increases. The consistent reduction in MSE with bin count empirically confirms that aggregating tokens with higher similarity directly minimizes the approximation error.

The single-bin formulation readily generalizes to multiple bins by partitioning $\mathcal{D}$ into $\{b_j\}_{j=1}^{N}$ and applying the same per-bin aggregation procedure.

4. Experiments

Empirical Validation of the *Low-Delta* Oracle. To verify the theoretical foundation, we evaluate the effect of grouping logits with higher similarity. Theorem 3.1 predicts that aggregating tokens with identical scores minimizes the approximation error. We focus on the discarded long-tail logits and apply our aggregation method while varying the number of discrete bins in powers of two from 1 to 32, and measure the MSE of the attention outputs across heads (Fig. 4). Although the absolute magnitude of the error and the impact of binning vary by head, a consistent trend is observed: as the number of bins increases, the MSE consistently decreases. Increasing the bin count reduces the in-bin variance of attention scores, so the aggregated tokens share a higher degree of similarity. This empirically validates the core theory underlying the proposed oracle policy: grouping tokens by score similarity reduces approximation error in real-world scenarios.

Effectiveness of UTA across Sparsity Ratios. To quantitatively verify the effectiveness of UTA, we measure the MSE of attention outputs against the full attention baseline using the Wan2.1-14B model (Wan et al., 2025). Our evaluation is twofold: (i) from an oracle perspective, we compare Top-K against Top-K+UTA; (ii) from an implemental application perspective, we integrate UTA into existing structured sparse attention frameworks, SVG2 (Yang et al., 2025a), under the premise that masked regions inherently correspond to *Low-Delta* indices. As shown in Table 1, integrating UTA consistently reduces the MSE across all sparsity settings and

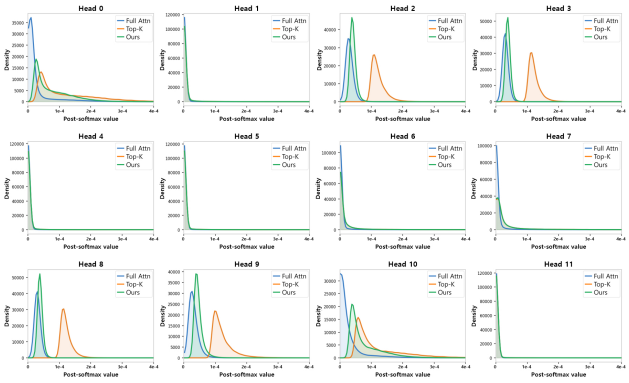


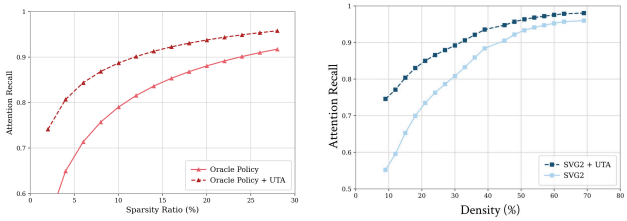
Figure 5. Density visualization of post-softmax values across individual attention heads. UTA (green) successfully approximates the full attention distribution (blue), preventing the probability inflation and rightward shift caused by Top-K truncation (orange).

baselines. From the oracle perspective at 50.0% sparsity, UTA reduces the MSE by approximately 97.40%, with a maximum error reduction of 8.06% when applied to SVG2. These findings demonstrate that UTA, grounded in the *Low-Delta* Oracle Policy, is a straightforward yet highly effective method for mitigating approximation errors in both Top-K approaches and existing sparse attention frameworks.

Preservation of Attention Distributions. To directly evaluate how well UTA maintains structural integrity, we visualize post-softmax value distributions across individual attention heads of Block 0 of Wan2.1-14B (Wan et al., 2025). As demonstrated in Fig. 5, our UTA consistently approximates the full attention probability density, while the Top-K baseline frequently exhibits a rightward shift. This effect is particularly pronounced in heads with widely spread distributions (Heads 2, 3, 8, 9): when the distribution is broad, the discarded continuous tail carries a substantial portion of the probability mass, and rigid truncation artificially inflates surviving probabilities. By unifying the discarded tail into a single representative and preserving its collective mass in the denominator, UTA inherently prevents this inflation. Two-dimensional attention map visualizations further corroborate this: Top-K produces severe noise and vertical artifacts in widely spread heads, whereas UTA yields clean attention maps visually indistinguishable from full attention.

Recall and Sparsity Analysis. Fig. 6 presents attention recall across varying sparsity levels from both the oracle perspective and the practical application perspective (SVG2). In both settings, integrating UTA yields consistent improvements in attention recall across all sparsity ratios, confirming that the *Low-Delta* policy captures more of the original attention mass than naive truncation.

Component Analysis. We compare two variants of UTA: recovering only the softmax normalization (denominator) versus recovering both normalization and attention value (numerator) terms. As shown in Fig. 7, the full UTA significantly reduces MSE across all sparsity ratios compared to the normalization-only ablation, confirming that both components are essential for minimizing structural approximation errors, consistent with the dual error decomposition in Eq. (6).



(a) Oracle Policy *(b)* Sparse Attention

Figure 6. Recall vs. sparsity ratio. UTA improves attention recall across both oracle and practical sparse attention settings.

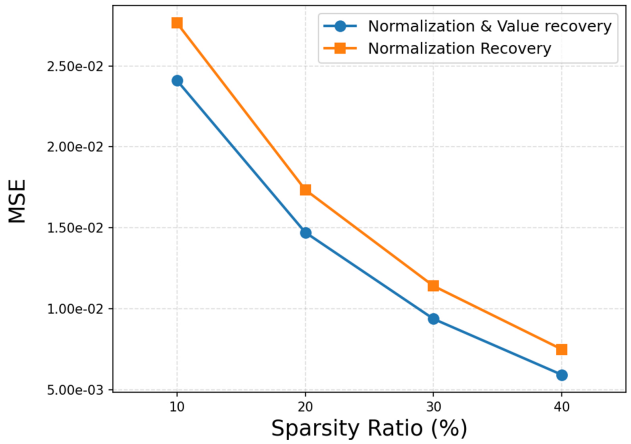


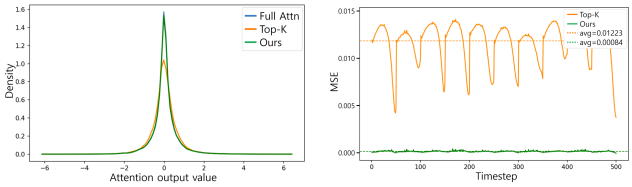
Figure 7. Component ablation of UTA: recovering only the normalization term (orange) vs. recovering both normalization and value terms (blue). Full UTA consistently achieves lower MSE.

Attention Output Distribution and Temporal Stability. Fig. 8(a) visualizes the overall attention output distribution. The Top-K baseline exhibits a flattened peak due to systemic renormalization distortion, whereas UTA flawlessly matches the full attention distribution. Fig. 8(b) evaluates the MSE across 500 diffusion timesteps at 25% sparsity. While Top-K suffers from high and fluctuating errors (average MSE 0.01223), UTA consistently maintains near-zero MSE (average 0.00084), an approximately 14.5× reduction, demonstrating strong temporal stability throughout the iterative denoising process.

Qualitative Results. Fig. 9 presents a qualitative comparison between baseline dense attention and Radial+UTA on Wan2.1-14B (Wan et al., 2025) across four scenarios. Our approach is robust under dynamic camera movement

Table 1. Quantitative comparison of MSE across different sparsity ratios on Wan2.1-14B. Integrating UTA significantly reduces approximation error not only for the Top-K baseline but also for recent sparse attention frameworks such as SVG2 (Yang et al., 2025a).

Method	**10.0%**	**20.0%**	**30.0%**	**40.0%**	**50.0%**
Top-K	6.03e-03	2.27e-03	1.02e-03	4.88e-04	2.32e-04
Top-K **+ UTA (Ours)**	**8.51e-04** (-85.89%)	**1.86e-04** (-91.81%)	**5.53e-05** (-94.58%)	**1.81e-05** (-96.29%)	**6.03e-06** (-97.40%)
SVG2	3.21e-02	1.57e-02	8.42e-03	4.42e-03	2.11e-03
SVG2 **+ UTA (Ours)**	**1.94e-02** (-39.56%)	**1.20e-02** (-23.57%)	**6.73e-03** (-20.07%)	**4.02e-03** (-9.05%)	**1.94e-03** (-8.06%)



(a) Output Distribution *(b)* MSE across Timesteps

Figure 8. Attention output analysis. **(a)** UTA preserves the output distribution while Top-K exhibits a flattened peak from systemic distortion. **(b)** UTA maintains near-zero MSE across 500 diffusion timesteps, whereas Top-K yields high and fluctuating errors.

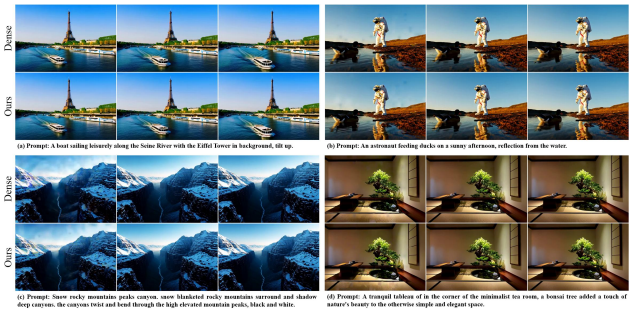


Figure 9. Visual comparison between dense attention and Radial+UTA on Wan2.1-14B (Wan et al., 2025) across (a) dynamic camera movement, (b) surreal object interactions, (c) complex natural landscapes, and (d) detailed indoor lighting.

and surreal object interactions while maintaining temporal consistency, and reconstructs complex natural landscapes and detailed indoor lighting with high-frequency details and subtle illumination preserved. The generated outputs are comparable to the dense baseline, supported quantitatively by a PSNR of 30.09, confirming that the efficient sparse attention mechanism effectively maintains the visual quality of the original dense generation.

5. Conclusion

In this work, we addressed the computational overhead of attention in video generation by identifying a fundamental flaw in Top-K and Top-P oracle policies: the rigid truncation discards the continuous tail and introduces compounding structural errors across iterative denoising steps. To resolve this, we introduced a novel *Low-Delta* Oracle Policy, mathematically proven to achieve zero error when identical attention scores are grouped (Theorem 3.1). As a practical implementation, we proposed Unified Tail Aggregation (UTA), which restores the full attention distribution by aggregating tail logits within a marginal delta into a single representative and compensating with $\log N_d$. Extensive evaluations demonstrated that UTA outperforms the Top-K oracle by up to 97.4% MSE reduction at 50% sparsity. Furthermore, integrating UTA into existing sparse attention frameworks (SVG2) significantly reduces the approximation error, confirming that masked regions in structured sparse methods inherently exhibit *Low-Delta* characteristics. Ultimately, by prioritizing the preservation of the entire attention distribution, our approach establishes a rigorous and highly effective foundation for scalable video generation.

Limitations and Future Work. The primary focus of this work is establishing a theoretical upper bound and a diagnostic oracle policy for sparse attention. Consequently, our evaluations assume an idealized oracle setting with prior knowledge of the full softmax distribution. Additionally, achieving optimal latency requires structural modifications to hardware-accelerated kernels such as FlashAttention (Zhang et al., 2025a), which is beyond our current scope. For future practical deployment, a promising direction is to approximate the required *Low-Delta* values via sub-sampling and natively integrate these operations into the inner loops of hardware-optimized kernels, realizing a highly efficient sparse attention mechanism.

Acknowledgements

This work was supported by Samsung Electronics Co., Ltd. (IO251218-14799-01). This research was supported by the IITP (Institute of Information Communications Technology Planning Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2026-RS-2023-00260091). In addition, this work was supported by the Institute of Information Communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-02263706, Development of an analog-digital mixed ultra-low power neuromorphic edge SoC).

References

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Pengtao Chen, Xianfang Zeng, Maosen Zhao, Peng Ye, Mingzhu Shen, Wei Cheng, Gang Yu, and Tao Chen. Sparse-vdit: Unleashing the power of sparse attention to accelerate video diffusion transformers. *arXiv preprint arXiv:2506.03065*, 2025a.

Ruichen Chen, Keith G Mills, Liyao Jiang, Chao Gao, and Di Niu. Re-ttention: Ultra sparse visual generation via attention statistical reshape. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Sankeerth Durvasula, Kavya Sreedhar, Zain Moustafa, Suraj Kothawade, Ashish Gondimalla, Suvinay Subramanian, Narges Shahidi, and Nandita Vijaykumar. Fg-attn: Leveraging fine-grained sparsity in diffusion transformers. *arXiv preprint arXiv:2509.16518*, 2025.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

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

Qirui Li, Guangcong Zheng, Qi Zhao, Jie Li, Bin Dong, Yiwu Yao, and Xi Li. Compact attention: Exploiting structured spatio-temporal sparsity for fast video generation. *arXiv preprint arXiv:2508.12969*, 2025a.

Xingyang Li, Muyang Li, Tianle Cai, Haocheng Xi, Shuo Yang, Yujun Lin, Lvmin Zhang, Songlin Yang, Jinbo Hu, Kelly Peng, et al. Radial attention: $O(n \log n)$ sparse attention with energy decay for long video generation. *arXiv preprint arXiv:2506.19852*, 2025b.

Akide Liu, Zeyu Zhang, Zhexin Li, Xuehai Bai, Yizeng Han, Jiasheng Tang, Yuanjie Xing, Jichao Wu, Mingyang Yang, Weihua Chen, et al. Fpsattention: Training-aware fp8 and sparsity co-design for fast video diffusion. *arXiv preprint arXiv:2506.04648*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

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

Xin Tan, Yuetao Chen, Yimin Jiang, Xing Chen, Kun Yan, Nan Duan, Yibo Zhu, Daxin Jiang, and Hong Xu. Dsv: Exploiting dynamic sparsity to accelerate large-scale video dit training. *arXiv preprint arXiv:2502.07590*, 2025.

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

Jianzong Wu, Liang Hou, Haotian Yang, Xin Tao, Ye Tian, Pengfei Wan, Di Zhang, and Yunhai Tong. Vmoba: Mixture-of-block attention for video diffusion models. *arXiv preprint arXiv:2506.23858*, 2025.

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

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

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15982–15993, 2025.

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

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

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Jintao Zhang, Haoxu Wang, Kai Jiang, Shuo Yang, Kaiwen Zheng, Haocheng Xi, Ziteng Wang, Hongzhou Zhu, Min Zhao, Ion Stoica, et al. Sla: Beyond sparsity in diffusion transformers via fine-tunable sparse-linear attention. *arXiv preprint arXiv:2509.24006*, 2025a.

Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. Spargeattn: Accurate sparse attention accelerating any model inference. *arXiv e-prints*, pp. arXiv–2502, 2025b.

Jintao Zhang, Kai Jiang, Chendong Xiang, Weiqi Feng, Yuezhou Hu, Haocheng Xi, Jianfei Chen, and Jun Zhu. Spargeattention2: Trainable sparse attention via hybrid top-k+ top-p masking and distillation fine-tuning. *arXiv preprint arXiv:2602.13515*, 2026.

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*, 2025c.

Yuechen Zhang, Jinbo Xing, Bin Xia, Shaoteng Liu, Bohao Peng, Xin Tao, Pengfei Wan, Eric Lo, and Jiaya Jia. Training-free efficient video generation via dynamic token carving. *arXiv preprint arXiv:2505.16864*, 2025d.

Tianchen Zhao, Ke Hong, Xinhao Yang, Xuefeng Xiao, Huixia Li, Feng Ling, Ruiqi Xie, Siqi Chen, Hongyu Zhu, Yichong Zhang, et al. Paroattention: Pattern-aware reordering for efficient sparse and quantized attention in visual generation models. *arXiv preprint arXiv:2506.16054*, 2025.