TOWARDS A LOW-DELTA ORACLE POLICY FOR IMPROVED UPPER BOUND IN SPARSE ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

While Diffusion Transformers (DiTs) deliver remarkable visual quality in video generation, the high cost of 3D spatio-temporal attention hinders their scalability. To evaluate and optimize sparse attention, existing studies rely on the Top-K Oracle Policy. However, the rigid truncation of this approach discards the continuous tail, introducing structural errors that accumulate during iterative diffusion. To address these limitations, we introduce a novel *Low-Delta* Oracle Policy based on a proof that sparse attention achieves zero error when grouping identical attention scores. Building on this, our oracle analysis shows that leveraging low-delta scores preserves the structural integrity of the distribution. Based on this observation, we propose a Unified Tail Aggregation (UTA) method, which aggregates logits where the score variance is bounded by a marginal delta, seamlessly integrating with efficient attention kernels while incurring less than 1% FLOP overhead in our evaluation. Extensive evaluations show our approach achieves up to a 97.4% MSE reduction and a 39.6% improvement under the Top-K oracle and an advanced sparse baseline, respectively.

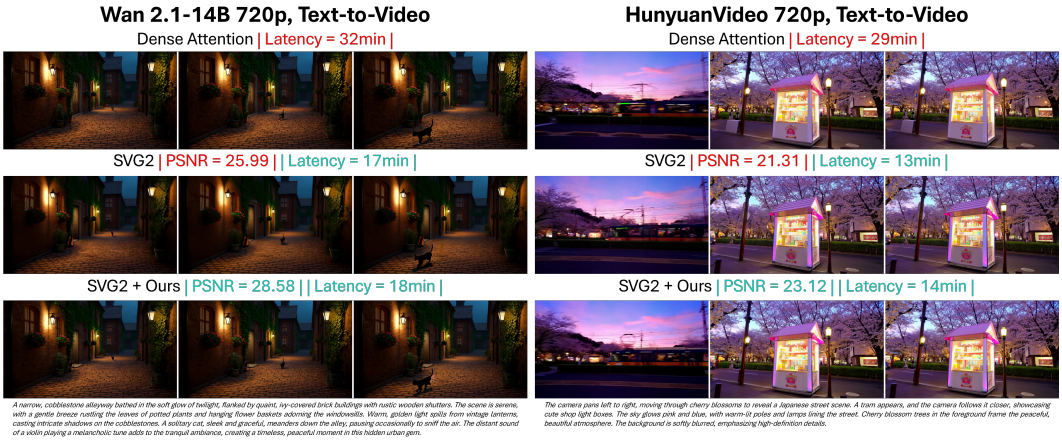


Figure 1: Our UTA effectively restores PSNR, improving by 2.59 (Wan 2.1) and 1.81 (HunyuanVideo), while maintaining the high acceleration efficiency of SVG2 on a single H100 GPU for HunyuanVideo and Wan2.1.

1 INTRODUCTION

Although 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; Chen et al., 2025a) introduces massive computational overhead. Relying on 3D spatio-temporal full attention yields high-quality consistent outputs, the exploding temporal token count renders it computationally prohibitive. To mitigate this, sparse attention mechanisms select critical key-value tokens for each query. While static methods (Xi et al., 2025; Zhang et al., 2025b; Li et al., 2025b) mitigate complexity through predetermined

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

attention masks, dynamic methods (Zhang et al., 2025a; Xu et al., 2025; Xia et al., 2025) seek to preserve relevant information by evaluating sparsity patterns during inference.

To establish a theoretical performance upper bound, existing studies predominantly rely on the Top-K oracle policy. Furthermore, several approaches (Zhu et al., 2025; Lin et al., 2025; Yang et al., 2025) adopt a Top-P oracle policy, which determines sparsity based on attention score values. However, we identify a limitation in this approach: as sparsity increases, the Top-K or Top-P method relies on rigid truncation that discards the continuous tail of the attention distribution. This rigid truncation distorts the inherent probability distribution of full attention. This distribution shift can create compounding errors across diffusion timesteps, severely degrading generation quality.

We argue that a true upper bound for sparse attention must fundamentally 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 have the exact same attention score, the sparse attention can be executed with zero error. Inspired by this theoretical foundation, we introduce *Low-Delta* Oracle Policy, a novel policy directed towards finding and exploiting logit indices with minimal differences in their attention scores.

To empirically verify our theoretical claims, we propose a Unified Tail Aggregation (UTA) method as a practical implementation of the *Low-Delta* Oracle Policy. UTA seamlessly integrates with efficient attention kernels such as FlashInfer (Ye et al., 2025) while incurring less than 1% FLOP overhead in our evaluation. By grouping tokens where the score variance is bounded by a marginal delta, UTA effectively compresses redundant contextual information while preserving the original attention distribution.

Furthermore, we evaluate our approach in both the Top-*K* oracle setting and a practical sparse attention setting using SVG2 (Yang et al., 2025). Experimental results demonstrate that UTA achieves up to a 97.4% reduction in MSE at a 50% density. Additionally, since the masked logits in SVG2 predominantly exhibit *Low-Delta* characteristics, applying UTA to these regions yields a 39.6% MSE reduction at a 10% density.

- We establish a zero-error condition for aggregating identical attention scores and introduce the *Low-Delta* Oracle Policy based on this principle to mitigate distribution distortion caused by rigid Top-K truncation.
- We propose Unified Tail Aggregation (UTA) as a practical implementation of the *Low-Delta* Oracle Policy. UTA restores the attention distribution by supplementing a single aggregated logit and seamlessly integrates with efficient attention kernels while incurring less than 1% FLOP overhead.
- We demonstrate that our approach outperforms in both oracle and practical settings. UTA achieves up to a 97.4% MSE reduction at a 50% density under the Top-*K* oracle setting, and a 39.6% improvement at an extreme 10% density when applied to masked regions in SVG2.

2 Related Works

2.1 Sparse Attention for Video Diffusion

Sparse attention for video diffusion has been actively researched to alleviate the excessive computational overhead of the attention mechanism. Static mask-based sparse attention methods (Xi et al., 2025; Zhang et al., 2025b; Li et al., 2025b;a; Chen et al., 2025b) apply predefined attention masks by analyzing the patterns of attention scores a priori. Conversely, dynamic methods (Zhang et al., 2025a; 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., 2025c; Wu et al., 2025; Durvasula et al., 2025; Xi et al., 2025; Yang et al., 2025; Li et al., 2026; Gu et al., 2025) apply sparsity masks by conducting small-scale analyses of the attention map's sparsity patterns during runtime. Both approaches conventionally adopt the Top-*K* selection of the attention map as their Oracle Policy. Recently, alternative methods have been introduced that determine the density based on probability mass using Top-*P* (Zhang et al., 2026; Yang et al., 2025; Zhu et al., 2025; Lin et al., 2025) rather than Top-*K*. However, these approaches fundamentally fail to resolve the problem of attention distribution shift caused by omitted logits when high sparsity is applied. While some works (Chen et al., 2025c; Yuan et al., 2024) introduce methods to cache and reuse masked logits, this incurs a critical trade-off with memory or computational overhead. To

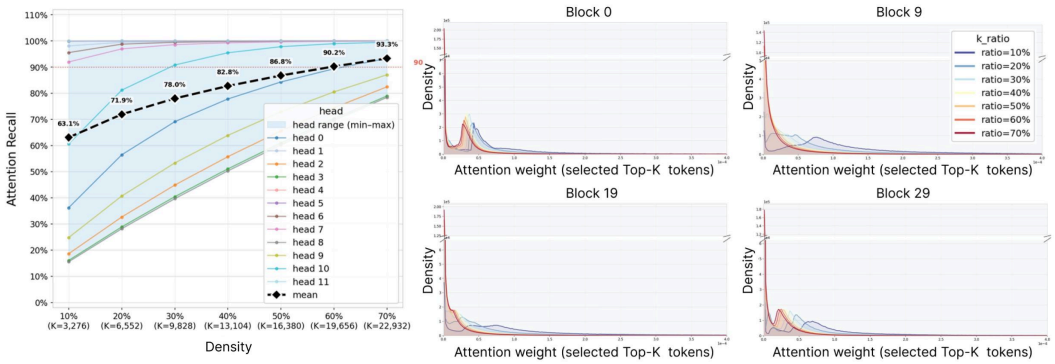


Figure 2: (a) Attention recall per head and (b) softmax distributions per block across varying density ratios. Notably, even in an oracle setting, the softmax distributions vary significantly with density.

address this, our study departs from the conventional Top-*K* or Top-*P* paradigms and presents a novel direction focused on identifying logits with similar attention scores.

3 METHOD

3.1 PRELIMINARIES

We first define full attention in matrix form before introducing the per-query expression. Given query $Q \in \mathbb{R}^{M\times d}$, key $K \in \mathbb{R}^{L\times d}$, and value $V \in \mathbb{R}^{L\times d}$, with M and L denoting the number of query and key-value tokens respectively, and d is the hidden dimension, the full attention output is defined as:

$$O \;=\; \mathrm{Attn}_{\mathrm{full}}(Q,K,V) \;\triangleq\; \mathrm{Softmax}\Big(\frac{QK^{\top}}{\sqrt{d}}\Big)\,V \tag{1}$$

For a single query token $\boldsymbol{q}_m$ (the mth row of Q), let $\mathcal{S}$ denote the index set of all tokens attending to $\boldsymbol{q}_m$, with corresponding key-value pairs $\{(\boldsymbol{k}_i, \boldsymbol{v}_i)\}_{i\in\mathcal{S}}$. The attention logit z_i and weight $p(i)$ for index $i \in \mathcal{S}$ are defined as:

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

Accordingly, the per-query attention output (the mth row of O) is written as:

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

3.2 THE FLAW OF TRUNCATION: DISTRIBUTION COLLAPSE IN TOP-K POLICY

To establish an upper bound for sparse attention, the prevailing standard across existing studies is the Top-K Oracle Policy. However, we argue that this naive truncation introduces a fundamental limitation that severely distorts the structural integrity of the attention, as shown in Fig. 2.

Let $\mathrm{Attn}_{\mathrm{full}}(\boldsymbol{q}_m)$ denote the full attention output for query $\boldsymbol{q}_m$ as defined in Eq. (3). Under rigid Top-K truncation, the full index set $\mathcal{S}$ is decomposed into two disjoint subsets, the selected set $\mathcal{T}$ and the discarded set $\mathcal{D}$:

$$\mathcal{S} = \mathcal{T} \sqcup \mathcal{D} \quad\Leftrightarrow\quad \mathcal{S} = \mathcal{T} \cup \mathcal{D},\;\; \mathcal{T} \cap \mathcal{D} = \varnothing, \tag{4}$$

where $|\mathcal{T}| = K$ and $\mathcal{D} = \mathcal{S} \setminus \mathcal{T}$.

Under Top-K truncation, the induced weight $p_{\mathcal{T}}(i)$ is the re-normalized distribution restricted to $\mathcal{T}$:

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

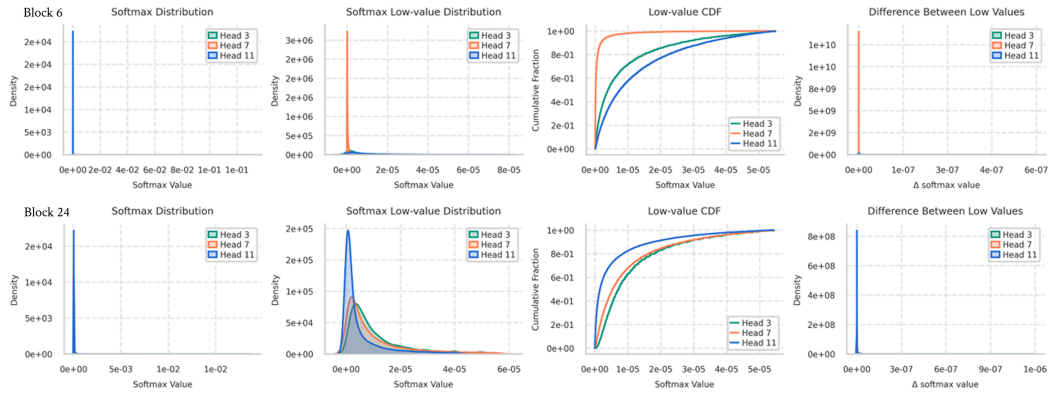


Figure 3: Analysis of the attention weight distributions. (a) Softmax distributions for full tokens. (b) Softmax distributions for lower-value tokens. (c) Cumulative Distribution Function (sCDF) for softmax values. (d) Distributions of the Δsoftmax values.

Here, $\alpha = \left(\sum_{j\in\mathcal{T}} p(j)\right)^{-1} \geq 1$ acts as an amplification factor that inflates all surviving weights to compensate for the removed probability mass, which directly induces the re-normalization distortion.

Using the per-query expression $\mathrm{Attn}_{\mathrm{full}}(\boldsymbol{q}_m) = \sum_{i\in\mathcal{S}} p(i)\boldsymbol{v}_i$ from Eq. (3) and the Top-K truncated output $\mathrm{Attn}_{\mathcal{T}}(\boldsymbol{q}_m) = \sum_{i\in\mathcal{T}} p_{\mathcal{T}}(i)\boldsymbol{v}_i$, the output discrepancy admits the following decomposition:

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

Eq. (6) shows that Top-K truncation distorts surviving weights and discards the tail, causing compounded errors across denoising steps in video diffusion models.

3.3 Theoretical Ideal for the Proposed *Low-Delta* Oracle Policy

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

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

where z_i is the attention logit defined in Eq. (2).

Theorem 1 (Zero-Error Aggregation)**.** *Let $\mathcal{S}$ be the full token set and $\mathcal{D} \subseteq \mathcal{S}$ be a truncation set, and let $N_d \triangleq |\mathcal{D}|$. For a given query token $\boldsymbol{q}_m$, assume that all attention logits over $\mathcal{D}$ are identical:*

$$z_i \triangleq \boldsymbol{q}_m^\top \boldsymbol{k}_i = \hat{z}_{\mathcal{D}}, \quad \forall i \in \mathcal{D}, \qquad \hat{z}_{\mathcal{D}} \triangleq \frac{1}{N_d}\sum_{i\in\mathcal{D}} z_i. \tag{8}$$

Define the aggregated value as $\hat{\boldsymbol{v}} \triangleq \frac{1}{N_d}\sum_{i\in\mathcal{D}} \boldsymbol{v}_i$. Then replacing the discarded logits $\{z_i\}_{i\in\mathcal{D}}$ and values $\{\boldsymbol{v}_i\}_{i\in\mathcal{D}}$ with a single representative pair $(\hat{z}_{\mathcal{D}}, \hat{\boldsymbol{v}})$, whose contribution is weighted by N_d, preserves the attention output exactly, i.e.,

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

and hence yields zero approximation error.

This theorem shows that aggregating tokens with identical logits can exactly preserve full attention outputs. The proof is provided in supplementary materials.

For non-identical tail logits, mean-logit aggregation incurs a Jensen gap in the tail partition mass. Appendix B bounds the resulting attention-output error, accounting for both logit and value variation. The bound vanishes for identical tail logits, recovering Theorem 1.

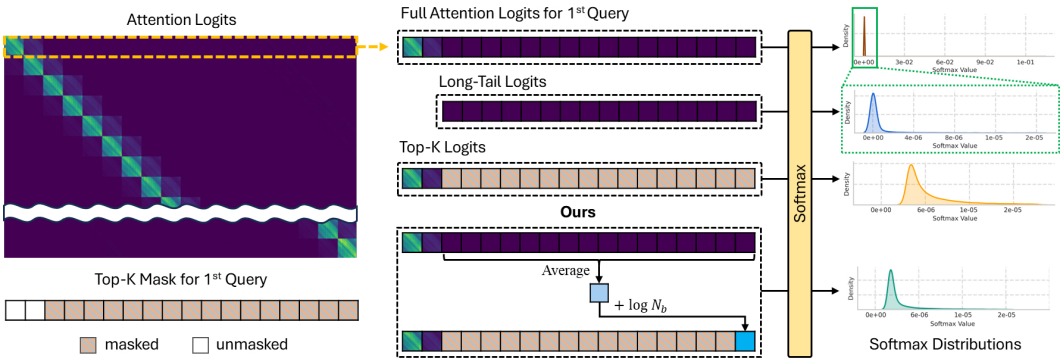


Figure 4: Overview of Unified Tail Aggregation (UTA) Framework: An illustration of our framework for the first query, simplified for clarity.

3.3.1 *Low-Delta* Oracle Policy.

Building on this ideal theory, the proposed *Low-Delta* oracle fundamentally shifts the objective of sparse attention. Instead of adhering to naive Top-K truncation, the *Low-Delta* oracle sets its primary goal toward identifying and exploiting indices that share similar attention scores. Grouping tokens into clusters based on similar attention score indices serves as a highly effective method to compress and utilize information while preserving the overall probability mass. Building on this principle, the following section introduces a simple procedure for implementing the *Low-Delta* oracle in practice.

3.4 Unified Tail Aggregation (UTA) based on *Low-Delta* Oracle Policy

3.4.1 Empirical Anatomy of Attention: Tail Logit Concentration.

To bridge the zero-error condition in Section 3.3 to practical implementations, we hypothesize that approximating tightly concentrated logits substantially minimizes error. We validate this by analyzing attention weight distributions in a standard video diffusion model. The score space is highly skewed; while a few tokens dominate, the lower 90% forms a dense cluster of small-magnitude logits near zero (Fig. 3(a, b)). Furthermore, differences between adjacent sorted values in this tail region form a sharp peak at zero (Fig. 3(d)), confirming they are nearly identical. These findings demonstrate that the tail is not mere noise but a dense, low-variance region.

3.4.2 Unified Tail Aggregation (UTA).

Motivated by the strong concentration of tail logits, as shown in Fig. 3, we present UTA in the simplest *single-bin* form, where the entire tail set $\mathcal{D}$ is aggregated into one representative token. An overview of the resulting UTA procedure is illustrated in Fig. 4 and detailed in Algorithm 1.

After obtaining the Top-K set $\mathcal{T}$, we aggregate the tail set $\mathcal{D}$ into a single bin $b \triangleq \mathcal{D}$ with size $N_d \triangleq |\mathcal{D}|$. Let the tail mean logit and mean value be

$$\bar{z}_{\mathcal{D}} \triangleq \frac{1}{N_d} \sum_{i \in \mathcal{D}} z_i, \qquad \hat{\boldsymbol{v}}_{\mathcal{D}} \triangleq \frac{1}{N_d} \sum_{i \in \mathcal{D}} \boldsymbol{v}_i, \tag{10}$$

where z_i is defined in Eq. (2). Then, the tail sums in the numerator and denominator of Eq. (7) admit the approximation

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

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

This approximation is justified by the logit concentration observed in Fig. 3, which implies $z_i \approx \bar{z}_{\mathcal{D}}$ for $i \in \mathcal{D}$. Following the zero-error aggregation principle, we compensate the aggregated tail logit:

$$\hat{z}_{\mathcal{D}} \triangleq \bar{z}_{\mathcal{D}} + \log N_d. \tag{13}$$

Finally, concatenating this single aggregated tail representative with the Top-K elements yields

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

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

3.4.3 Implementation: Integrating UTA with Hardware Efficient Kernel

Our approach is designed to merge the UTA proxy directly with the output of FlashInfer (Ye et al., 2025), introducing less than 1% FLOP overhead, as reported in Table 2. Algorithm 1 provides a practical implementation mathematically equivalent to Section 3.4.2.

We first compute the partial attention output ($O_{\mathcal{T}}$) and Log-Sum-Exp ($LSE_{\mathcal{T}}$) for the active Top-K clusters ($\mathcal{T}$), while representing the masked tail set ($\mathcal{D}$) as a single proxy token with its mean value ($\hat{\boldsymbol{v}}_{\mathcal{D}}$) and a log-based compensated logit ($\hat{z}_{\mathcal{D}}$). Because FlashInfer inherently returns $LSE_{\mathcal{T}}$ in a base-2 logarithm, we first apply a base conversion, $LSE_{\mathcal{T},\ln} = LSE_{\mathcal{T}} \times \ln(2)$, to properly align the scales. To yield the final aggregated output without recomputing the Softmax over $\mathcal{T}$, we then merge $O_{\mathcal{T}}$ and $\hat{\boldsymbol{v}}_{\mathcal{D}}$ using the standard numerically stable Softmax merging logic. By defining $m = \max(LSE_{\mathcal{T},\ln}, \hat{z}_{\mathcal{D}})$, we calculate the blending weights:

$$\begin{aligned} w_{\mathcal{T}} &= \exp(LSE_{\mathcal{T},\ln} - m), \\ w_{\mathcal{D}} &= \exp(\hat{z}_{\mathcal{D}} - m), \\ O_{\text{final}} &= \frac{w_{\mathcal{T}} O_{\mathcal{T}} + w_{\mathcal{D}} \hat{\boldsymbol{v}}_{\mathcal{D}}}{w_{\mathcal{T}} + w_{\mathcal{D}}} \end{aligned} \tag{15}$$

By factoring out the scaling constant $\exp(-m)$, the exponential shift cancels out, verifying that $\exp(LSE_{\mathcal{T},\ln}) = \sum_{n\in\mathcal{T}} \exp(z_n)$ and $\exp(LSE_{\mathcal{T},\ln}) O_{\mathcal{T}} = \sum_{n\in\mathcal{T}} \exp(z_n)\boldsymbol{v}_n$. Therefore, this hardware-efficient merging matches the theoretical UTA formulation described in Eq. (14), ensuring zero theoretical deviation.

Algorithm 1 Unified Tail Aggregation

Require: $Q, K, V \in \mathbb{R}^{S\times D}$, $\bar{K}, \bar{V} \in \mathbb{R}^{K_b\times D}$
Require: $M \in \{\text{True}, \text{False}\}^{Q_b\times K_b}$
Require: $W \in \mathbb{R}^{K_b}, \mathcal{B}_b$
Ensure: Out $\in \mathbb{R}^{S\times D}$

$O, LSE \leftarrow \text{FlashInfer}(Q, K, V, M)$
$LSE \leftarrow LSE \cdot \ln(2)$
for each block $b \in \{1, \ldots, Q_b\}$ **do**
  for each query $i \in \mathcal{B}_b$ **do**
    $m_i \leftarrow LSE_i, \quad l_i \leftarrow 1.0$
    $\text{acc}_i \leftarrow O_i$
    for each j s.t. $M_{b,j} = \text{False}$ **do**
      $S_{i,j} \leftarrow \frac{Q_i \bar{K}_j^\top}{\sqrt{D}} + \ln(W_j)$
      $m_{\text{new}} \leftarrow \max(m_i, S_{i,j})$
      $\alpha \leftarrow \exp(m_i - m_{\text{new}})$
      $p \leftarrow \exp(S_{i,j} - m_{\text{new}})$
      $l_i \leftarrow l_i \cdot \alpha + p$
      $\text{acc}_i \leftarrow \text{acc}_i \cdot \alpha + p\bar{V}_j$
      $m_i \leftarrow m_{\text{new}}$
    end for
    $\text{Out}_i \leftarrow \text{acc}_i / l_i$
  end for
end for
return Out

4 Experiments

4.1 Empirical Validation of Low-Delta Oracle Policy

To verify the theoretical foundation of our Oracle Policy, we first conducted an analysis to evaluate the effect of grouping logits with higher similarity. According to our theoretical zero error condition, aggregating tokens with identical scores minimizes the approximation error. To test this principle in an practical setting, we focused on the discarded long tail logits and applied our aggregation method. In Fig. 5, we varied the number of discrete bins in powers of two from 1 to 32 and measured the MSE of the attention outputs across different heads. The absolute magnitude of the error and the specific impact of binning vary depending on the individual

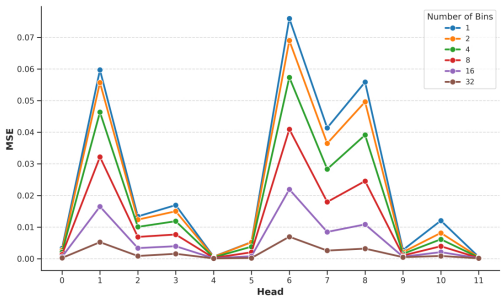


Figure 5: MSE of the attention outputs across heads with varying numbers of bins.

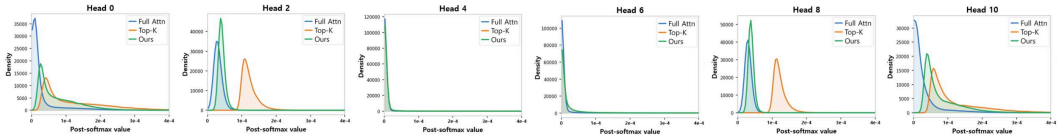


Figure 6: Density visualization of post-softmax values across individual attention heads in Wan 2.1-14B model (Wan et al., 2025). The proposed method (green) successfully approximates the full attention distribution (blue), actively preventing the probability inflation and rightward shift typically caused by the Top-K truncation (orange).

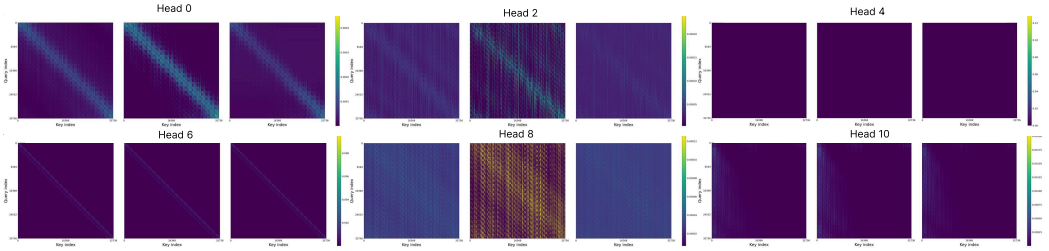


Figure 7: Visual comparison of 2D attention maps across heads, demonstrating that UTA preserves full attention patterns without the severe noise and artifacts introduced by the Top-K baseline.

head. However, a consistent trend is clearly observable across all evaluated heads: as the number of bins increases, the MSE consistently decreases. As the number of bins increases, the score variance within each bin decreases, so the aggregated tokens become more similar.

4.2 Preservation of Attention Distributions via UTA

Softmax Distribution. To directly evaluate how well UTA maintains the structural integrity of the attention mechanism, we visualized the post-softmax value distributions across individual attention heads. As demonstrated in Fig. 6, our UTA method consistently approximates the full attention probability density across all evaluated heads. In contrast, the Top-K baseline frequently exhibits a rightward shift, failing to capture the true probability landscape. This distribution-preserving effect of UTA is particularly pronounced in heads where the attention distribution is relatively spread out over a wider range of values, such as Heads 0, 2, 8, and 10. When a head exhibits a wider distribution, the discarded continuous tail contains a substantial portion of the total probability mass. Rigidly truncating this heavy tail in the Top-K approach induces a significant normalization error that artificially inflates the probabilities of the surviving tokens.

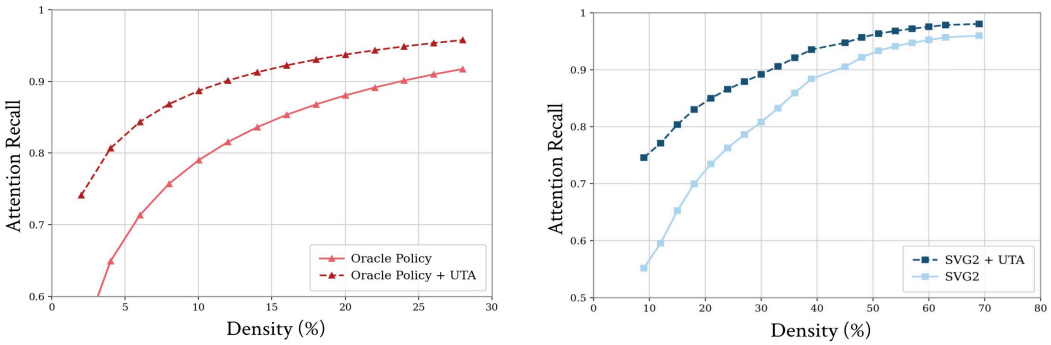
Attention Map. To further contextualize the impact of our head-wise distribution analysis, Fig. 7 presents 2D attention maps comparing the full attention, the Top-K baseline, and our UTA method. Consistent with the probability density results in Fig. 6, the Top-K approach corrupts structural patterns of the attention matrix, particularly in heads characterized by widely spread distributions such as Heads 2 and 8. In these specific heads, the artificial probability inflation caused by rigid truncation visibly manifests as severe noise and vertical artifacts, completely destroying the full attention structure. Conversely, our UTA method successfully compensates for this truncation error, yielding clean attention maps that are visually indistinguishable from the full attention.

4.3 Effectiveness of Unified Tail Aggregation

To quantitatively verify the effectiveness of the proposed UTA across varying density levels, we measure the MSE of the attention outputs compared to the full attention baseline using the Wan 2.1-14B model (Wan et al., 2025). In Table 1, our evaluation is twofold: from an Oracle Policy perspective, we compare the Top-K oracle against Top-K enhanced with UTA, and from an implemental application perspective, we integrate UTA into an existing structured sparse attention framework, specifically SVG2 (Yang et al., 2025). For the implemental application, we apply UTA under the premise that the masked regions in SVG2 inherently correspond to *Low-Delta* indices. As detailed in Table 1, integrating UTA consistently reduces the MSE across all evaluated density settings and baseline

Table 1: Quantitative comparison of MSE across different density levels. Integrating UTA significantly reduces approximation errors for both Top-K and SVG2 (Yang et al., 2025).

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) Oracle Policy (b) Sparse Attention

Figure 8: Comparison of Attention Recall across density levels for oracle and sparse attention policies.

methods. Notably, UTA achieves maximum MSE reductions of 97.40% and 39.56% under the Oracle Policy and SVG2 (Yang et al., 2025), respectively. Furthermore, as shown in Fig. 8, we conducted the analysis of the attention recall across varying density levels. From both the oracle policy perspective and the application perspective, the integration of UTA yields significant improvements in attention recall across all evaluated density settings. Additional comparisons with SVG, Radial attention, DFSAttn, SVOO, and sliding-window attention are provided in Appendix C.7.

4.4 Quality and Efficiency Benchmark on Video Generation

For our video-generation experiments, we use the VBench text-to-video prompt suite to generate approximately five-second videos at 720p resolution with Wan 2.1-14B and HunyuanVideo-13B.

To verify how restoring the discarded attention distribution impacts video generation quality, we specifically integrated UTA under the highly sparse SVG2-Turbo condition. As reported in Table 2, we evaluate video generation quality using PSNR, SSIM, LPIPS (Zhang et al., 2018), and VBench (Huang et al., 2024).

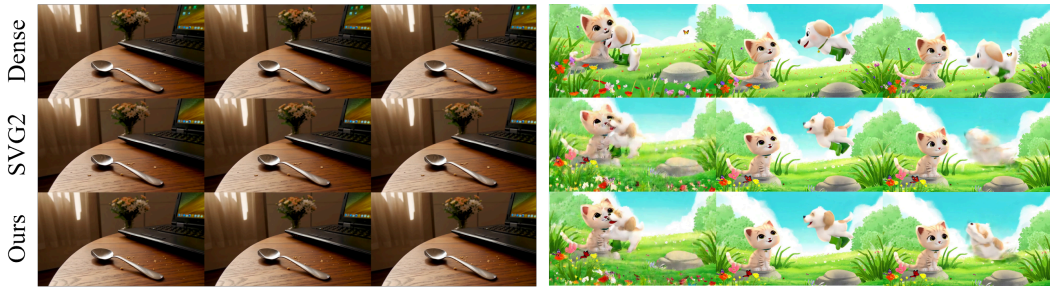
We compare each generated video with its full-attention counterpart for PSNR, SSIM, and LPIPS. VBench provides a reference-free assessment of video quality. Following SVG2 (Yang et al., 2025), we report VBench scores as the arithmetic mean over five dimensions: subject consistency, background consistency, motion smoothness, aesthetic quality, and imaging quality. Scores are expressed as percentages. Compared to SVG2-Turbo, our method achieves higher PSNR (24.032) and SSIM (0.823), alongside a lower LPIPS (0.136). Furthermore, on Wan 2.1-14B, UTA improves the VBench score from 83.8% to 84.5%, approaching the full-attention baseline of 84.6%. We additionally report FVD against the full-attention video distribution in Appendix C.8, where our method achieves lower values than SVG2-Turbo on both models. Integrating UTA increases the FLOP count by less than 1% relative to SVG2-Turbo, while slightly reducing inference speedup. Specifically, the speedup relative to full attention decreases from 1.77× to 1.66× on Wan 2.1-14B and from 1.97× to 1.91× on HunyuanVideo.

4.5 Qualitative Results

As illustrated in Fig. 9, we present a qualitative comparison between the baseline dense attention and our proposed method, which integrates SVG2 (Yang et al., 2025) with UTA. While the standalone SVG2 framework accelerates generation, its rigid truncation under high sparsity inevitably leads

Table 2: Quality and efficiency benchmarking results on Wan 2.1-14B and Hunyuan, Text-to-Video.

Method	PSNR↑	SSIM↑	LPIPS↓	VBench↑	Density↓	FLOP↓	Speedup↑
Wan 2.1-14B (720P)							
Full Attention	–	–	–	84.6	100%	658.46 PFLOPs	1×
SpargeAttn	20.519	0.623	0.343	82.0	42.03%	468.46 PFLOPs	1.41×
SVG	22.989	0.785	0.199	83.7	30.25%	429.86 PFLOPs	1.53×
SVG2-Turbo	23.682	0.789	0.196	83.8	12.87%	372.89 PFLOPs	1.77×
Ours	24.032	0.823	0.136	84.5	13.02%	375.14 PFLOPs	1.66×
Hunyuan-13B (720p)							
Full Attention	–	–	–	84.2	100%	612.38 PFLOPs	1.00×
SVG2-Turbo	27.709	0.846	0.231	82.2	13.71%	291.74 PFLOPs	1.97×
Ours	28.552	0.859	0.214	82.2	13.67%	292.93 PFLOPs	1.91×



(a) Prompt: a spoon and a laptop. **(b) Prompt: a cat and a dog.**

Figure 9: Qualitative results of generated videos on Wan 2.1-14B. Comparison of video frames produced by Dense, SVG2, and Ours across different prompts.

to noticeable visual degradation and structural artifacts. Consequently, our approach restores high visual fidelity, achieving generation quality that closely matches the original dense attention baseline. Additional qualitative results are provided in Appendix D.

5 Conclusion

In this work, we addressed the computational overhead of 3D spatio-temporal 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. To resolve this, we introduced a novel *Low-Delta* Oracle Policy, mathematically proven to achieve zero error when identical attention scores are aggregated. As a practical implementation of the *Low-Delta* oracle, we proposed the Unified Tail Aggregation (UTA) method, which seamlessly integrates with FlashInfer while incurring less than 1% FLOP overhead. Extensive evaluations demonstrated that UTA consistently outperforms the Top-K oracle. Furthermore, integrating UTA into an existing sparse attention method significantly reduces the approximation error, demonstrating improvements in video generation quality metrics, including PSNR and VBench.

5.1 Limitations and Future Work

Although UTA is technically compatible with generative models such as Wan and Hunyuan and with FlashAttention/FlashInfer-based sparse frameworks for T2V and I2V tasks, GPU constraints prevented exhaustive evaluation across all model–task combinations. Its effectiveness also depends on whether the masked regions produced by the underlying sparse method capture low-delta logits. UTA may not reduce approximation error for every attention head (Appendix C.9), and lower attention-output MSE does not necessarily imply better perceptual quality for individual videos. Future work will broaden validation and develop UTA-specific sparse masks.

AI USE STATEMENT

In this work, we used generative AI tools (LLM-based assistants) to assist with translating our drafts into English. We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for or write the proofs, propose or refine hypotheses, design or provide feedback on the research methodology or experiments, implement the methods, or interpret the results, and generating synthetic datasets, cleaning or reformatting datasets, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to search for and identify related literature, to edit the text for readability, to format the LaTeX source, tables, captions, and references, and to check the consistency of the numbers reported across the text, tables, and appendices. We have reviewed all AI-assisted work: the authors read every AI-suggested reference in its original source before citing it, checked every translated or AI-edited passage against its intended meaning, and verified every reported number against our experiment logs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, user studies, or personal data. All models (Wan 2.1-14B and HunyuanVideo-13B), sparse attention baselines, and the VBench prompt suite are publicly available and used under their licenses. UTA is an inference-time approximation that adds no new generative capability; the generated content, including any biases, is determined by the underlying pretrained models. However, by lowering the cost of high-resolution video generation, efficiency methods such as ours can also make misleading synthetic media, such as deepfakes, cheaper to produce, so accelerated generators should be deployed with safeguards such as content filtering and watermarking. Like any approximate attention method, UTA can change model outputs and may increase the error for a small fraction of heads (Appendix C.9), so output quality should be validated on the target model and task before deployment.

REPRODUCIBILITY STATEMENT

The attention formulation and the Top-K error decomposition are given in Section 3.1 and Eq. (6). The zero-error condition of the *Low-Delta* oracle policy is stated in Theorem 1 and proved in Appendix A, and the approximation bound for non-identical tail logits, together with its assumptions, is given in Appendix B. UTA is defined in Section 3.4.2 (Eqs. (10)–(14)), and its FlashInfer-based implementation is given in Algorithm 1. The setup and results of the attention-level evaluation are described in Section 4 (Table 1 and Figs. 5 and 8) and Appendix C.7, and those of the video generation benchmark, including the models, prompts, and metrics, in Section 4 (Table 2) and Appendix C.8. The setup of the per-head error analysis is given in Appendix C.9. All models, baselines, and benchmarks we use are publicly available, and code will be released with the final version.

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.

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, et al. Sana-video: Efficient video generation with block linear diffusion transformer. *arXiv preprint arXiv:2509.24695*, 2025a.

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*, 2025b.

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*, 2025c.

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.

Youping Gu, Xiaolong Li, Yuhao Hu, Minqi Chen, and Bohan Zhuang. BLADE: Block-sparse attention meets step distillation for efficient video generation. *arXiv preprint arXiv:2508.10774*, 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.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 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.

Shenggui Li, Runyu Lu, Qiaoling Chen, Haiyan Yin, Yueming Lyu, Yonggang Wen, Ivor Tsang, and Tianwei Zhang. DSA: Efficient inference for video generation models via distributed sparse attention. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

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.

Chaofan Lin, Jiaming Tang, Shuo Yang, Hanshuo Wang, Tian Tang, Boyu Tian, Ion Stoica, Song Han, and Mingyu Gao. Twilight: Adaptive attention sparsity with hierarchical top-p pruning. *arXiv preprint arXiv:2502.02770*, 2025.

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*, 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. 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*, 2025.

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.

Zihao Ye, Lequn Chen, Ruihang Lai, Wuwei Lin, Yineng Zhang, Stephanie Wang, Tianqi Chen, Baris Kasikci, Vinod Grover, Arvind Krishnamurthy, and Luis Ceze. Flashinfer: Efficient and customizable attention engine for llm inference serving. *arXiv preprint arXiv:2501.01005*, 2025.

Zhihang Yuan, Hanling Zhang, Pu Lu, Xuefei Ning, Linfeng Zhang, Tianchen Zhao, Shengen Yan, Guohao Dai, and Yu Wang. Ditfastattn: Attention compression for diffusion transformer models. *Advances in Neural Information Processing Systems*, 37:1196–1219, 2024.

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, 2025a.

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*, 2025b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

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*, 2025c.

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.

Kan Zhu, Tian Tang, Qinyu Xu, Yile Gu, Zhichen Zeng, Rohan Kadekodi, Liangyu Zhao, Ang Li, Arvind Krishnamurthy, and Baris Kasikci. Tactic: Adaptive sparse attention with clustering and distribution fitting for long-context llms. *arXiv preprint arXiv:2502.12216*, 2025.

A Proof: Zero-Error Aggregation

Theorem 2 (Zero-Error Aggregation). *Let $\mathcal{S}$ be the full token set and $\mathcal{D} \subseteq \mathcal{S}$ be a truncation set, and let $N_d \triangleq |\mathcal{D}|$. For a given query token $\boldsymbol{q}_m$, assume that all attention logits over $\mathcal{D}$ are identical:*

$$z_i \triangleq \boldsymbol{q}_m^\top \boldsymbol{k}_i = \hat{z}_\mathcal{D}, \quad \forall i \in \mathcal{D}, \qquad \hat{z}_\mathcal{D} \triangleq \frac{1}{N_d} \sum_{i \in \mathcal{D}} z_i. \tag{16}$$

We define the aggregated value as $\hat{\boldsymbol{v}} \triangleq \frac{1}{N_d} \sum_{i \in \mathcal{D}} \boldsymbol{v}_i$. Then replacing the discarded logits $\{z_i\}_{i \in \mathcal{D}}$ and values $\{\boldsymbol{v}_i\}_{i \in \mathcal{D}}$ with a single representative pair $(\hat{z}_\mathcal{D}, \hat{\boldsymbol{v}})$, whose contribution is weighted by N_d, preserves the attention output exactly, i.e.,

$$\mathrm{Attn}_{\mathrm{dense}}(\boldsymbol{q}_m) \;=\; \mathrm{Attn}_{\mathrm{agg}}(\boldsymbol{q}_m) = \frac{\sum_{n \in \mathcal{T}} \exp(z_n) \boldsymbol{v}_n + N_d \exp(\hat{z}_\mathcal{D}) \hat{\boldsymbol{v}}}{\sum_{n \in \mathcal{T}} \exp(z_n) + N_d \exp(\hat{z}_\mathcal{D})} \tag{17}$$

and hence yields zero approximation error.

Proof. We first express the dense attention output for query $\boldsymbol{q}_m$ over the full token set $\mathcal{S} = \mathcal{T} \cup \mathcal{D}$, where $\mathcal{T}$ denotes the selected token set, $\mathcal{D}$ denotes the truncation set, and $\mathcal{T} \cap \mathcal{D} = \varnothing$:

$$\mathrm{Attn}_{\mathrm{dense}}(\boldsymbol{q}_m) \triangleq \frac{\sum_{i \in \mathcal{S}} \exp(z_i) \boldsymbol{v}_i}{\sum_{j \in \mathcal{S}} \exp(z_j)} = \frac{\sum_{n \in \mathcal{T}} \exp(z_n) \boldsymbol{v}_n + \sum_{i \in \mathcal{D}} \exp(z_i) \boldsymbol{v}_i}{\sum_{n \in \mathcal{T}} \exp(z_n) + \sum_{i \in \mathcal{D}} \exp(z_i)}, \tag{18}$$

where $z_i \triangleq \boldsymbol{q}_m^\top \boldsymbol{k}_i$ denotes the attention logit for token i, and $\boldsymbol{v}_i$ is its corresponding value vector.

Let $N_d \triangleq |\mathcal{D}|$ be the number of tokens in $\mathcal{D}$, and define the aggregated value as

$$\hat{\boldsymbol{v}} \triangleq \frac{1}{N_d} \sum_{i \in \mathcal{D}} \boldsymbol{v}_i. \tag{19}$$

Under the assumption that all logits in $\mathcal{D}$ are identical, we have

$$z_i = \hat{z}_\mathcal{D}, \qquad \forall i \in \mathcal{D}, \qquad \hat{z}_\mathcal{D} \triangleq \frac{1}{N_d} \sum_{i \in \mathcal{D}} z_i. \tag{20}$$

Therefore, the $\mathcal{D}$-dependent terms in Eq. (18) can be simplified as

$$\sum_{i \in \mathcal{D}} \exp(z_i) = N_d \exp(\hat{z}_\mathcal{D}), \tag{21}$$

$$\sum_{i \in \mathcal{D}} \exp(z_i) \boldsymbol{v}_i = \exp(\hat{z}_\mathcal{D}) \sum_{i \in \mathcal{D}} \boldsymbol{v}_i = N_d \exp(\hat{z}_\mathcal{D}) \hat{\boldsymbol{v}}. \tag{22}$$

Substituting these into Eq. (18) yields

$$\mathrm{Attn}_{\mathrm{dense}}(\boldsymbol{q}_m) = \frac{\sum_{n \in \mathcal{T}} \exp(z_n) \boldsymbol{v}_n + N_d \exp(\hat{z}_\mathcal{D}) \hat{\boldsymbol{v}}}{\sum_{n \in \mathcal{T}} \exp(z_n) + N_d \exp(\hat{z}_\mathcal{D})} = \mathrm{Attn}_{\mathrm{agg}}(\boldsymbol{q}_m). \tag{23}$$

Hence, the aggregated form preserves the dense attention output exactly, which proves zero approximation error. □

B Approximation Error for Non-Identical Tail Logits

We analyze the single-proxy aggregation in Eq. (14) for a fixed query and a fixed partition of valid keys into a retained set $\mathcal{T}$ and a nonempty tail set $\mathcal{D}$. All logarithms are natural, and the analysis assumes exact arithmetic.

Setup. Let $N_d = |\mathcal{D}|$ and define

$$\begin{aligned} \bar{z} &= \frac{1}{N_d}\sum_{i\in\mathcal{D}} z_i, \quad \bar{\boldsymbol{v}} = \frac{1}{N_d}\sum_{i\in\mathcal{D}} \boldsymbol{v}_i, \\ \delta_i &= z_i - \bar{z}, \qquad \Delta = \max_{i\in\mathcal{D}} |\delta_i|. \end{aligned} \tag{24}$$

Thus, $\sum_{i\in\mathcal{D}} \delta_i = 0$ and $\sum_{i\in\mathcal{D}} (\boldsymbol{v}_i - \bar{\boldsymbol{v}}) = 0$.

Define the retained contributions and tail statistics as

$$\begin{aligned} A &= \sum_{j\in\mathcal{T}} e^{z_j}, & \boldsymbol{B} &= \sum_{j\in\mathcal{T}} e^{z_j}\boldsymbol{v}_j, \\ a &= N_d e^{\bar{z}}, & m &= \frac{1}{N_d}\sum_{i\in\mathcal{D}} e^{\delta_i}, \\ \boldsymbol{r} &= \frac{1}{N_d}\sum_{i\in\mathcal{D}} e^{\delta_i}(\boldsymbol{v}_i - \bar{\boldsymbol{v}}). \end{aligned} \tag{25}$$

Here, a is the proxy tail mass, while the exact tail mass is am. The vector $\boldsymbol{r}$ captures the coupling between exponential logit weights and centered values.

Jensen gap. The log-domain gap between the exact and proxy tail masses is

$$g_J = \log\left(\sum_{i\in\mathcal{D}} e^{z_i}\right) - (\bar{z} + \log N_d) = \log m \geq 0. \tag{26}$$

Indeed, Jensen's inequality gives

$$m = \frac{1}{N_d}\sum_{i\in\mathcal{D}} e^{\delta_i} \geq \exp\left(\frac{1}{N_d}\sum_{i\in\mathcal{D}} \delta_i\right) = 1. \tag{27}$$

Consequently, mean-logit aggregation underestimates the tail mass by $a(m-1)$, where $m - 1 = e^{g_J} - 1$.

Exact error decomposition. The full-attention and single-proxy outputs can be written as

$$\boldsymbol{O}_{\text{full}} = \frac{\boldsymbol{B} + am\bar{\boldsymbol{v}} + a\boldsymbol{r}}{A + am}, \qquad \boldsymbol{O}_{\text{agg}} = \frac{\boldsymbol{B} + a\bar{\boldsymbol{v}}}{A + a}. \tag{28}$$

Using $\boldsymbol{B} = (A+a)\boldsymbol{O}_{\text{agg}} - a\bar{\boldsymbol{v}}$ yields

$$\boldsymbol{O}_{\text{full}} - \boldsymbol{O}_{\text{agg}} = \frac{a}{A + am}\left[\boldsymbol{r} + (m-1)\left(\bar{\boldsymbol{v}} - \boldsymbol{O}_{\text{agg}}\right)\right]. \tag{29}$$

The first term captures the error from replacing exponentially weighted tail values with their arithmetic mean. The second captures the output error associated with the tail-mass approximation.

Approximation bound. Define

$$\begin{aligned} \sigma_z^2 &= \frac{1}{N_d}\sum_{i\in\mathcal{D}} \delta_i^2, \quad \sigma_v^2 = \frac{1}{N_d}\sum_{i\in\mathcal{D}} \|\boldsymbol{v}_i - \bar{\boldsymbol{v}}\|_2^2, \\ W &= \frac{a}{A+a}, \qquad D_V = \max_{i,j\in\mathcal{T}\cup\mathcal{D}} \|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2. \end{aligned} \tag{30}$$

Here, W is the normalized proxy mass; the exact normalized tail mass is $am/(A+am)$. The value diameter D_V is taken over both retained and tail keys.

Theorem 3 (Approximation bound for non-identical tail logits)**.** *Under the definitions above,*

$$\|\boldsymbol{O}_{\text{full}} - \boldsymbol{O}_{\text{agg}}\|_2 \leq e^{\Delta} W\left[\sigma_z\sigma_v + \frac{1}{2}(1-W)\sigma_z^2 D_V\right]. \tag{31}$$

Proof. Since the centered values sum to zero,

$$\boldsymbol{r} = \frac{1}{N_d} \sum_{i \in \mathcal{D}} (e^{\delta_i} - 1)(\boldsymbol{v}_i - \bar{\boldsymbol{v}}). \tag{32}$$

For $|\delta_i| \le \Delta$, the mean value theorem gives $|e^{\delta_i} - 1| \le e^{\Delta} |\delta_i|$. The triangle and Cauchy–Schwarz inequalities therefore yield

$$\begin{aligned} \|\boldsymbol{r}\|_2 &\le \frac{e^{\Delta}}{N_d} \sum_{i \in \mathcal{D}} |\delta_i| \, \|\boldsymbol{v}_i - \bar{\boldsymbol{v}}\|_2 \\ &\le e^{\Delta} \sigma_z \sigma_v. \end{aligned} \tag{33}$$

Next, the second-order Taylor remainder satisfies

$$0 \le e^{\delta_i} - 1 - \delta_i \le \frac{1}{2} e^{\Delta} \delta_i^2. \tag{34}$$

Averaging over $\mathcal{D}$ and using $\sum_{i \in \mathcal{D}} \delta_i = 0$ gives

$$0 \le m - 1 \le \frac{1}{2} e^{\Delta} \sigma_z^2. \tag{35}$$

In particular, the log-domain Jensen gap satisfies $0 \le g_J = \log m \le m - 1$.

Since $\bar{\boldsymbol{v}}$ is a convex combination of tail values, its distance to any retained value is at most D_V. Hence,

$$\begin{aligned} \|\bar{\boldsymbol{v}} - \boldsymbol{O}_{\text{agg}}\|_2 &= \left\| \frac{\sum_{j \in \mathcal{T}} e^{z_j} (\bar{\boldsymbol{v}} - \boldsymbol{v}_j)}{A + a} \right\|_2 \\ &\le \frac{A}{A + a} D_V = (1 - W) D_V. \end{aligned} \tag{36}$$

This also holds when $\mathcal{T}$ is empty, in which case $A = 0$, $W = 1$, and $\boldsymbol{O}_{\text{agg}} = \bar{\boldsymbol{v}}$.

Finally, $m \ge 1$ implies $a/(A + am) \le W$. Applying the preceding inequalities to Eq. (29) gives

$$\begin{aligned} \|\boldsymbol{O}_{\text{full}} - \boldsymbol{O}_{\text{agg}}\|_2 &\le W \left[\|\boldsymbol{r}\|_2 + (m - 1)(1 - W) D_V \right] \\ &\le e^{\Delta} W \left[\sigma_z \sigma_v + \frac{1}{2} (1 - W) \sigma_z^2 D_V \right], \end{aligned} \tag{37}$$

which proves the bound. □

The bound separates logit–value coupling from the tail-mass approximation error. When all tail logits are identical, $\sigma_z = 0$, $m = 1$, and $\boldsymbol{r} = \boldsymbol{0}$, so the exact zero-error result is recovered. For non-identical tails, the bound quantifies the joint dependence on logit spread, value variation, and proxy mass for the fixed query and partition. If the tail is empty, no proxy is introduced and the aggregation error is zero.

C Additional Experimental Results

C.1 Attention Weight Distributions

In Fig. 10, we present additional visualization results of the attention weight distributions across various heads for three representative blocks (Blocks 6, 12, and 24). Consistent with our main findings, the overall softmax distributions (Fig. 10a) and the zoomed-in lower-value distributions (Fig. 10b) demonstrate a strong long-tailed property, where the dense regions of the attention weights are tightly clustered near zero. The Cumulative Distribution Function (CDF) plots (Fig. 10c) further corroborate this observation, showing a steep convergence at extremely small values across all observed heads, regardless of the block depth. Furthermore, as illustrated in Fig. 10d, the differences between adjacent lower-values (Δ softmax value) are strictly concentrated at zero.

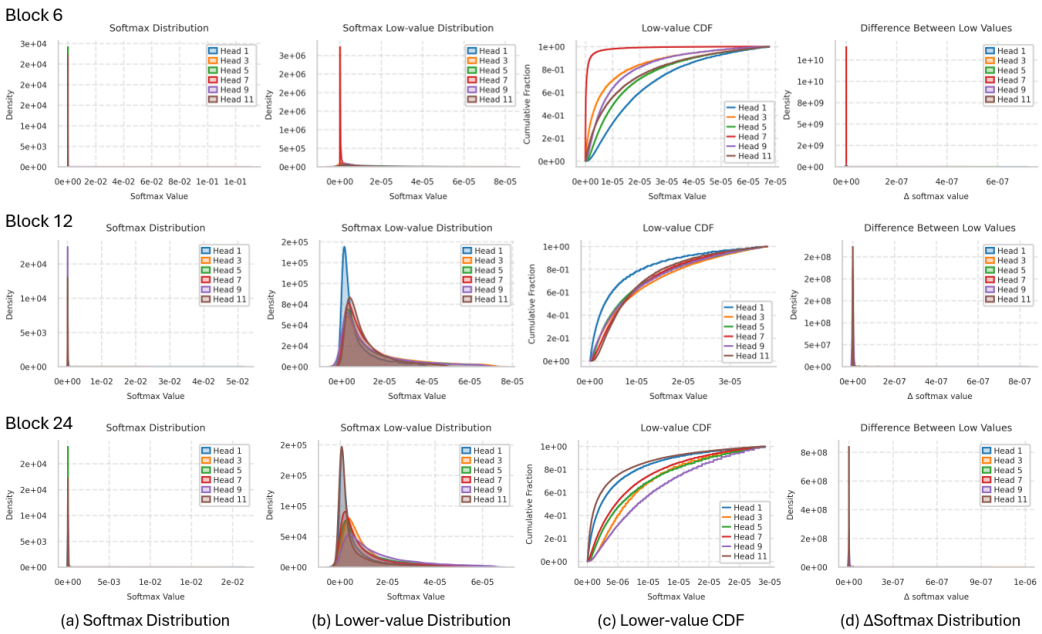


Figure 10: Additional visualization of the attention weight distributions. (a) Softmax distributions for full tokens. (b) Softmax distributions for lower-value tokens. (c) Cumulative Distribution Function (CDF) for softmax values. (d) Distributions of the Δsoftmax values.

C.2 Post-Softmax Distribution Compensation

In Fig. 11, we provide additional density visualization results of the post-softmax values across individual attention heads. We specifically selected Blocks 2, 5, and 29 to demonstrate the robustness and effectiveness of our approach under conditions where the attention scores exhibit a more dispersed (or widely distributed) pattern. As illustrated in the figure, in heads where the attention weights are not strictly concentrated at a single point, the conventional Top-K approach significantly distorts the original density distribution. This distortion occurs because the naive Top-K method drops the tail values without proper compensation, leading to an overestimation of the remaining post-softmax values. Conversely, our proposed approach closely aligns with the original Dense Attention distribution (blue shaded area) across all examined heads. These results clearly demonstrate that our method is particularly powerful when attention scores are highly dispersed, effectively preserving the structural integrity of the attention matrix by accurately compensating for the truncation errors.

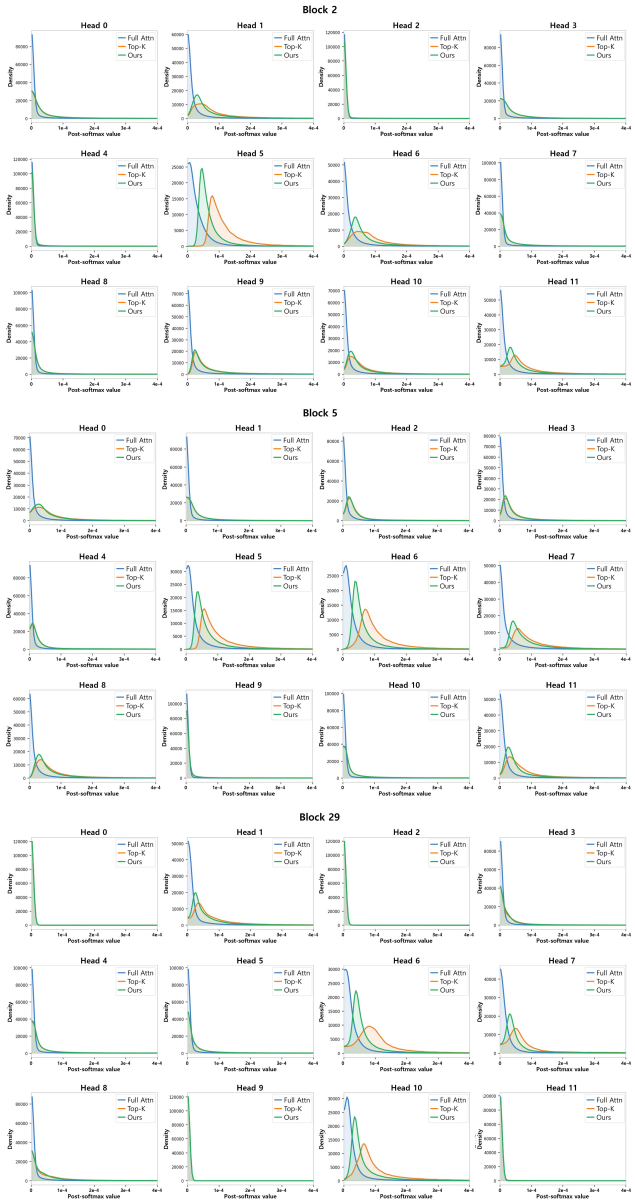


Figure 11: Additional density visualization of post-softmax values across individual attention heads.

C.3 Attention Map Compensation

In Fig. 12, we provide an additional visual comparison of the 2D attention maps across individual heads, specifically focusing on Block 1. This visualization clearly illustrates how our approach compensates for the structural distortions introduced by the naive Top-K method. As shown in the figure, the conventional Top-K approach tends to severely distort the original structural patterns of the attention matrix. Particularly in heads with more dispersed or complex attention patterns (e.g., Heads 2, 3, and 8), Top-K artificially amplifies the selected sparse scores without proper normalization compensation, resulting in noisy and over-activated artifacts compared to the Dense Attention baseline. In contrast, our approach effectively compensates for the truncation errors of the discarded lower values. The resulting attention maps of our method are visually almost indistinguishable from the original Dense Attention maps across all heads. These qualitative results further confirm that our approach successfully preserves the structural integrity of the dense attention matrix, even in the early layers (Block 1) where attention patterns are often highly globally distributed.

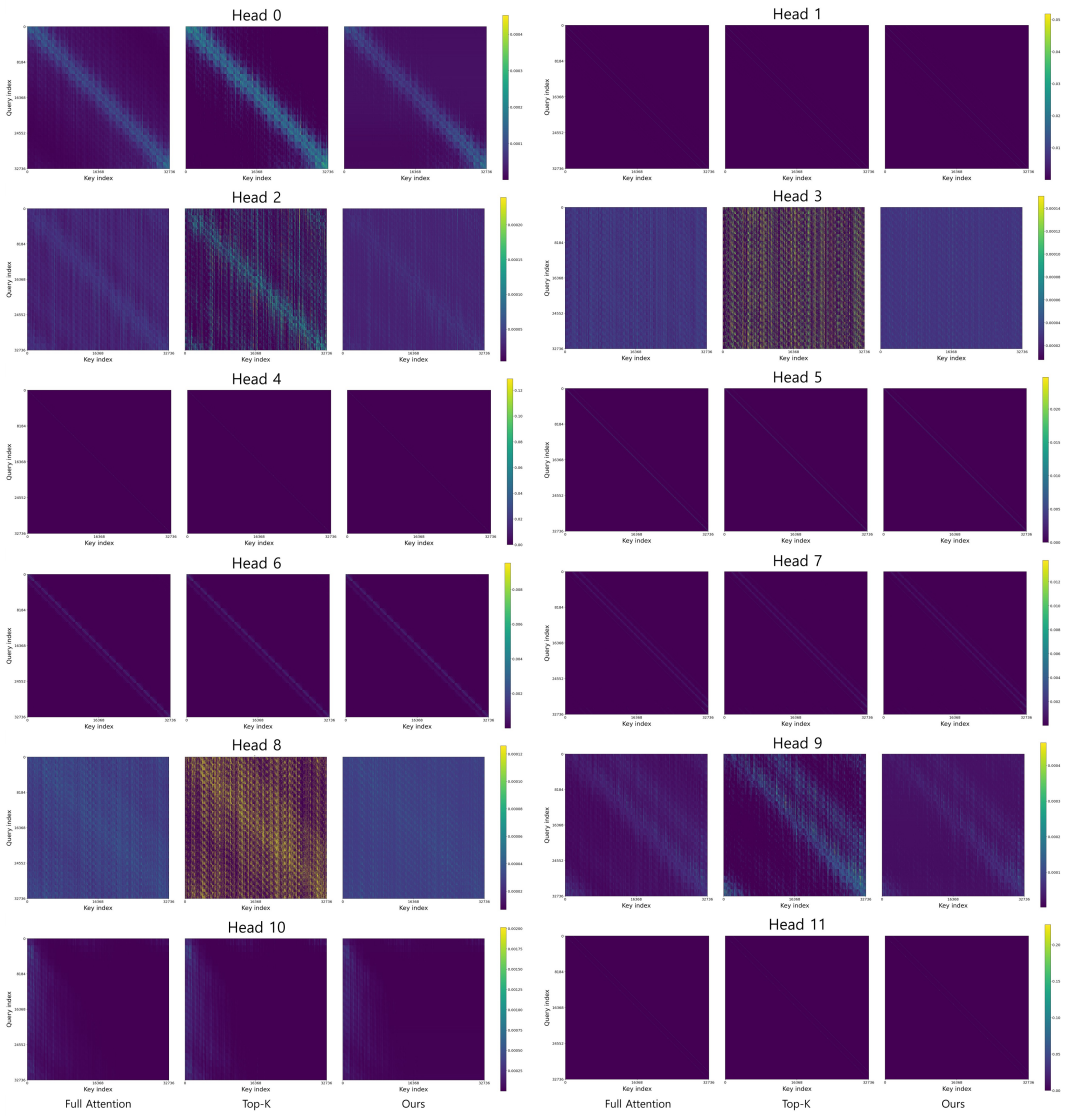
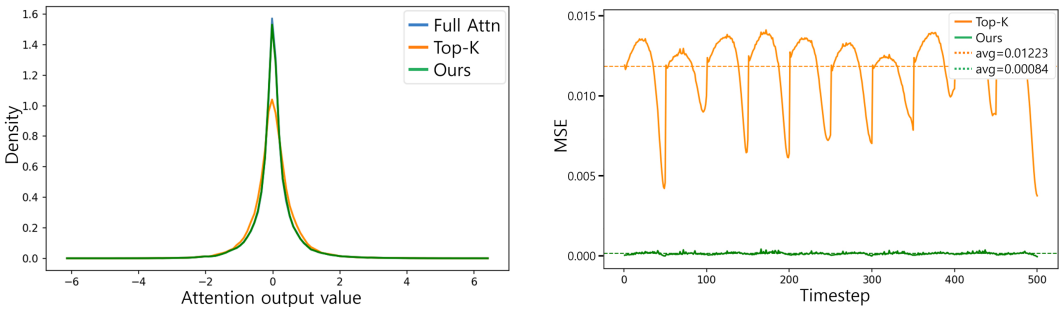


Figure 12: Additional visual comparison of 2D attention maps across individual heads.

C.4 Attention Output Distribution and Timestep Error

As shown in Fig. 13 (a), we visualized the attention output distribution to analyze its structural integrity. The experimental results demonstrate that while the Top-K approach exhibits a flattened peak and noticeably deviates from the full attention baseline, our proposed method perfectly preserves the original distribution. This near identical alignment with the full attention density explicitly confirms that our approach effectively prevents the systemic distortion caused by naive truncation. As shown in Fig. 13 (b), we conducted an experiment at a 25% density to analyze the MSE of the Top-K method and our proposed *Low-Delta* oracle across diffusion timesteps. The experimental results demonstrate that while the existing Top-K approach yields an average MSE of 0.01223, our proposed method achieves a significantly lower average MSE of 0.00084. This indicates a remarkable improvement, reducing the approximation error by approximately 14.5 times on average across the entire timestep range.



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

Figure 13: Analysis of the final attention outputs. **(a)** The overall probability density of the attention output values. The Top-K baseline exhibits a flattened peak due to systemic distortion, whereas our UTA method flawlessly matches the Full Attention distribution. **(b)** MSE evaluated across 500 diffusion timesteps. While Top-K suffers from high and fluctuating errors, UTA consistently maintains near-zero MSE, proving its temporal stability.

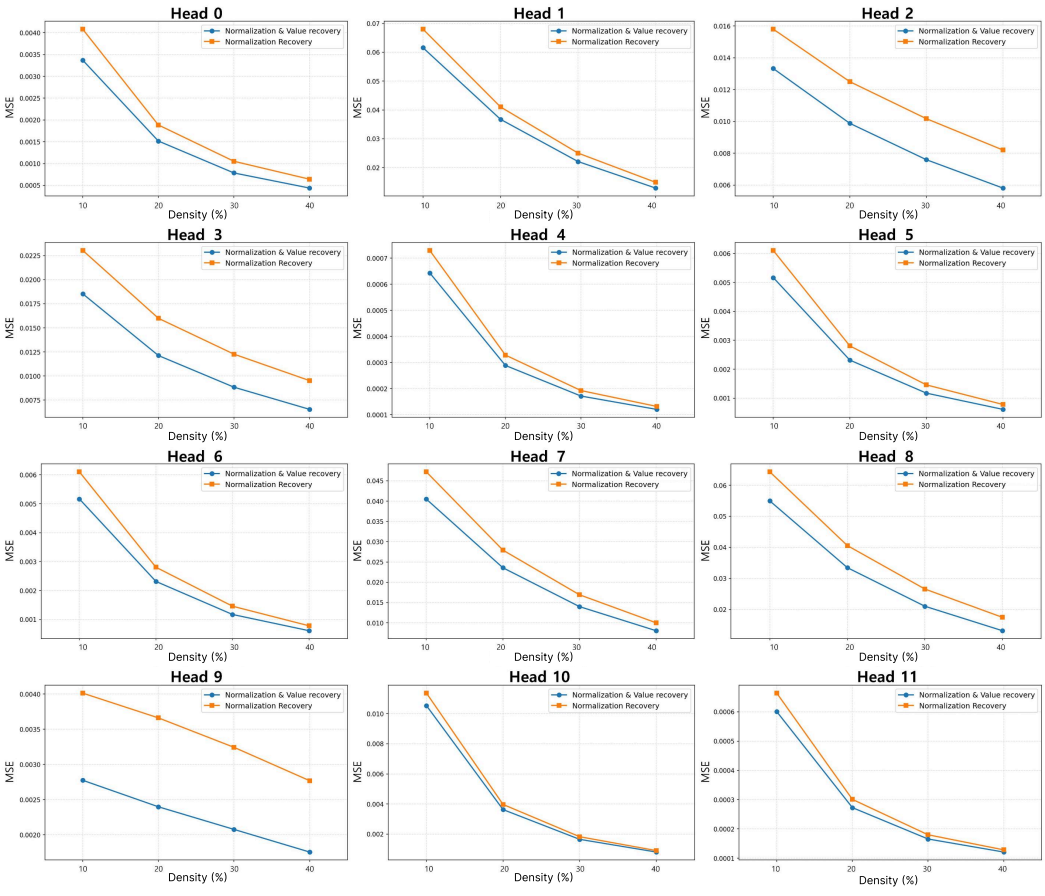


Figure 14: Additional ablation results on components of UTA for individual heads by comparing when recovering only the softmax normalization term versus our full approach that recovers both normalization and attention value terms across density ratios.

C.5 Ablation: Normalization vs. Full Recovery

In Fig. 14, we present additional ablation results to thoroughly validate the individual components of our Unified Tail Aggregation (UTA) approach. Specifically, we compare the Mean Squared Error (MSE) of the approximated attention outputs across individual heads under various density ratios. The comparison is drawn between an incomplete baseline that recovers only the softmax normalization term (orange line) and our full UTA approach that jointly recovers both the normalization and the attention value terms (blue line). As illustrated in the figure, our full approach consistently achieves a lower MSE across all evaluated heads compared to the normalization-only baseline. This indicates that solely compensating for the denominator of the softmax operation is insufficient; the discarded tail values hold meaningful structural information that must be aggregated into the final output.

C.6 Effect of UTA Across Network Depths

Table 3 summarizes the attention-output MSE recovery of UTA relative to SVG2 across early, middle, and late layer groups. Wan 2.1-14B exhibits no monotonic trend with depth, whereas HunyuanVideo shows greater recovery in later layers. Both models exhibit positive recovery at the reported group level.

Table 3: Attention-output MSE recovery of UTA relative to SVG2 across network depths. Layer ranges are shown in parentheses. Recovery values are expressed as percentages.

Layer group	Wan 2.1-14B	HunyuanVideo
Early	51.1 ± 6.0 (L0–9)	32.9 ± 15.4 (L0–19)
Middle	35.0 ± 7.4 (L10–29)	36.8 ± 13.7 (L20–39)
Late	52.9 ± 10.4 (L30–39)	69.3 ± 15.9 (L40–59)

C.7 UTA Across Different Sparse Attention Baselines

To examine how the effectiveness of UTA varies with the underlying sparse attention method, we compare attention-output MSE for SVG2, SVG, Radial attention, DFSAttn, SVOO, and sliding-window attention, with and without UTA. As shown in Table 4, UTA reduces the reported MSE in every available method–density pair, while the magnitude of improvement varies across methods. These results extend the attention-level evaluation beyond SVG2.

Table 4: Attention-output MSE relative to full attention for different sparse attention methods, with and without UTA. Columns indicate attention density. Lower values are better; "–" denotes an unreported entry.

Method	10%	20%	30%	40%	50%
SVG2	3.21e-02	1.57e-02	8.42e-03	4.42e-03	2.11e-03
SVG2 + UTA	1.94e-02	1.20e-02	6.73e-03	4.02e-03	1.94e-03
SVG	4.09e-02	1.71e-02	8.18e-03	4.48e-03	3.09e-03
SVG + UTA	2.69e-02	1.24e-02	6.14e-03	3.45e-03	2.49e-03
Radial	–	–	4.58e-03	3.14e-03	2.05e-03
Radial + UTA	–	–	2.98e-03	2.04e-03	1.34e-03
DFSAttn	1.33e-02	6.89e-03	3.95e-03	2.32e-03	1.35e-03
DFSAttn + UTA	7.13e-03	3.98e-03	2.42e-03	1.48e-03	8.84e-04
SVOO	1.10e-02	5.07e-03	2.74e-03	1.55e-03	8.72e-04
SVOO + UTA	2.43e-03	8.44e-04	3.80e-04	1.91e-04	9.98e-05
Sliding Window	1.85e-02	1.17e-02	8.15e-03	6.08e-03	4.57e-03
Sliding Window + UTA	1.27e-02	8.27e-03	5.80e-03	4.42e-03	3.41e-03

C.8 FVD RESULTS

To evaluate distribution-level fidelity, we compute FVD using videos generated with full attention on each backbone as the model-specific reference distribution.

Table 5: FVD against the full-attention reference for each model. Lower is better.

Model	SVG2-Turbo	Ours	Reduction (%)↑
Wan 2.1-14B (720p)	35.71	**31.29**	**12.4**
HunyuanVideo-13B (720p)	92.16	**90.53**	**1.8**

As shown in Table 5, our method reduces FVD from 35.71 to 31.29 on Wan 2.1-14B and from 92.16 to 90.53 on HunyuanVideo-13B, corresponding to relative reductions of 12.4% and 1.8%, respectively. The consistent improvement indicates that UTA brings the generated-video distribution closer to the full-attention reference on both models, with a larger gain on Wan 2.1-14B.

C.9 HEAD-LEVEL ERROR ANALYSIS OF UTA

We conduct a per-head error analysis of UTA on Wan and HunyuanVideo using 10 prompts. We examine how often UTA increases the approximation error relative to the sparse baseline and quantify the magnitude of these increases. Each head-level observation corresponds to one attention head evaluated in an attention call. The reported counts aggregate observations across the evaluated configurations, prompts, layers, and timesteps. The per-head analysis excludes layer 0, which is designated for full attention in the evaluation configuration.

Evaluation metrics. Let e_{S} and e_{UTA} denote the per-head MSEs of the sparse baseline and UTA, respectively, using the full-attention output as the reference. We define the recovery ratio as

$$R = \frac{e_{\mathrm{S}} - e_{\mathrm{UTA}}}{e_{\mathrm{S}}}.$$

A positive recovery ratio indicates that UTA reduces the approximation error. We classify an observation as adverse when $e_{\mathrm{UTA}} > e_{\mathrm{S}}$. For these observations, we measure the relative MSE increase as

$$I = \frac{e_{\mathrm{UTA}} - e_{\mathrm{S}}}{e_{\mathrm{S}}}.$$

The adverse rate is calculated over all evaluated observations, while the median and 95th percentile of I are calculated only over adverse observations.

Table 6: Per-head error analysis of UTA. Adverse rates use all evaluated observations as the denominator, while MSE increase statistics describe only the adverse subset. The analysis uses timestep values $0 < t \leq 958$ for Wan and $0 < t \leq 968.114929$ for HunyuanVideo.

Metric	Wan	HunyuanVideo
Head-level observations	1,966,760	1,131,384
Adverse observations	3,824	10,149
Adverse rate (%)	0.19	0.90
Median recovery (%)	+31.01	+23.45
Adverse increase: median (%)	0.0245	30.04
Adverse increase: 95th pct. (%)	1.23	565.70

Results and interpretation. As shown in Table 6, UTA achieves positive median recovery ratios of 31.01% on Wan and 23.45% on HunyuanVideo. Adverse observations account for 0.19% and 0.90% of the respective evaluation subsets. On Wan, the MSE increases within the adverse subset are generally small, with a median of 0.0245% and a 95th percentile of 1.23%. On HunyuanVideo, the corresponding increases reach 30.04% and 565.70%, showing that infrequent adverse cases can still involve substantial local errors.

Although individual heads can exhibit substantial error increases, adverse observations account for less than 1% of the evaluated cases on both models. Together with the positive median recovery, these results show that UTA improves approximation accuracy for the large majority of evaluated head-level observations.

D Additional Qualitative Results

In Figs. 15 and 16, we additionally provide more diverse qualitative results of the baseline dense attention, the Top-K approach and our approach. Furthermore, we provide additional qualitative comparisons in Figs. 17 and 18 to evaluate our method against SVG2 (Yang et al., 2025). These results show that our approach more accurately preserves the visual quality of the original dense generation compared to the Top-K and SVG2 baselines.

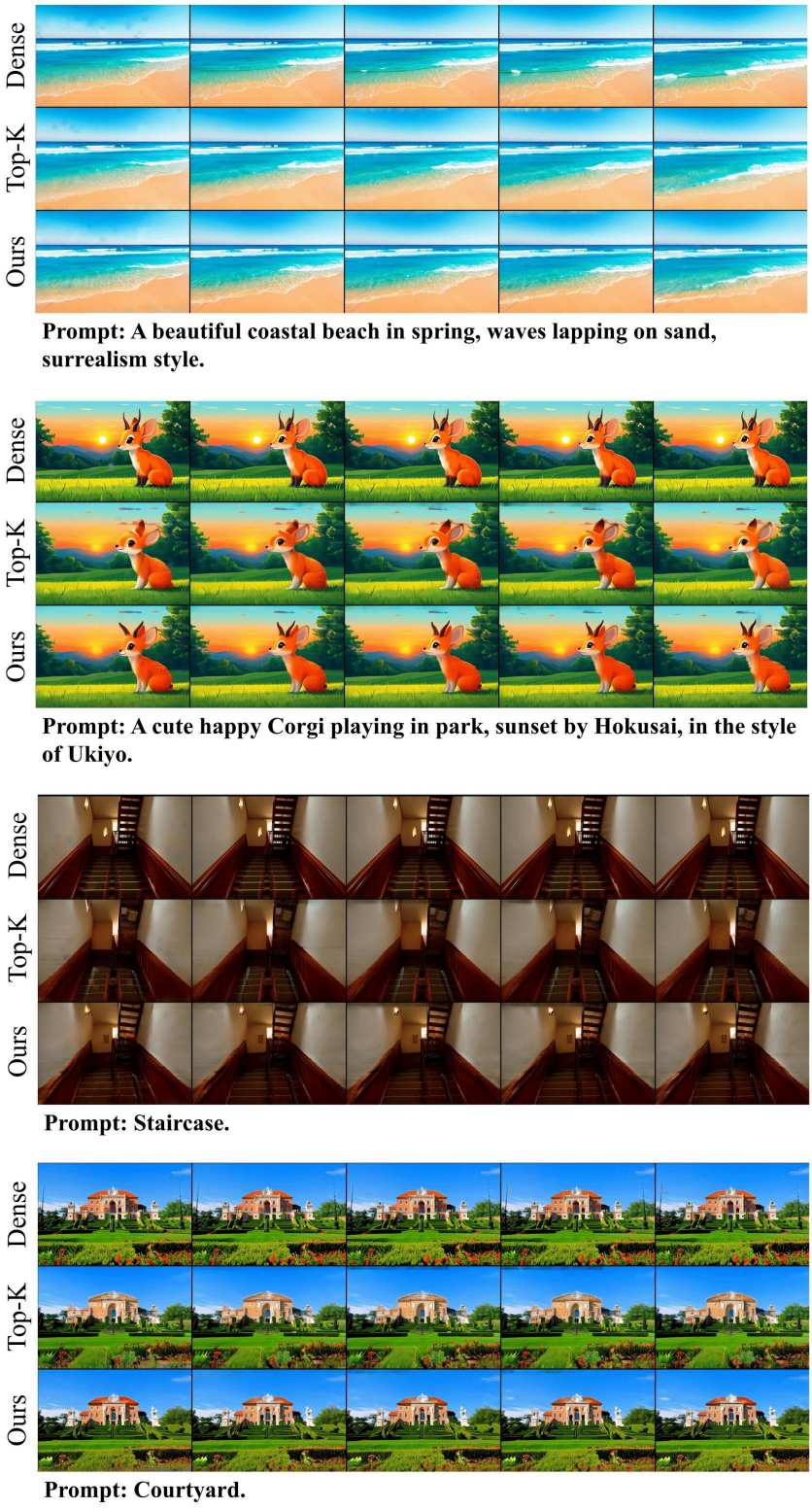


Figure 15: Additional qualitative comparisons of the dense attention, the Top-K baseline and our method.

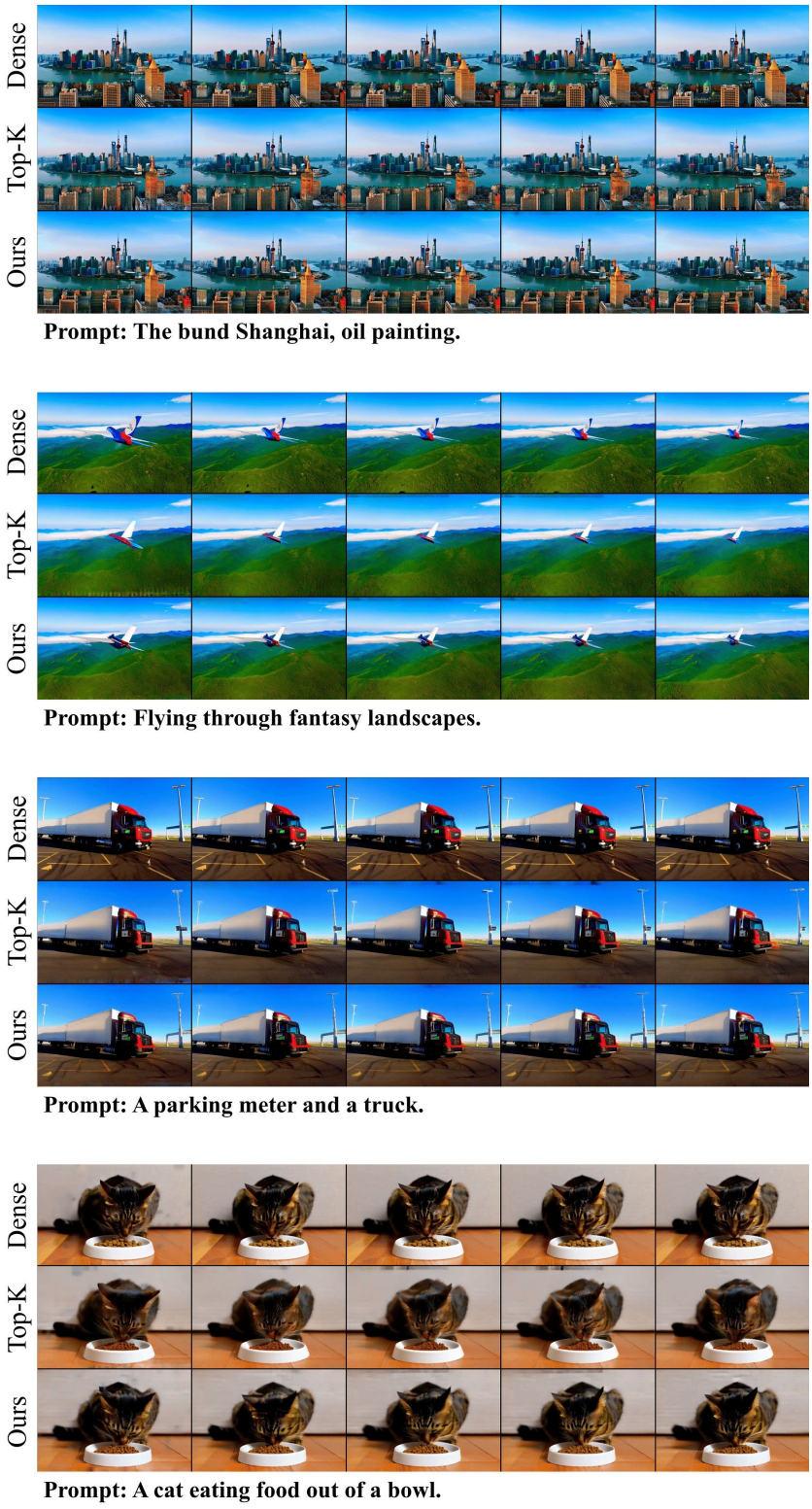
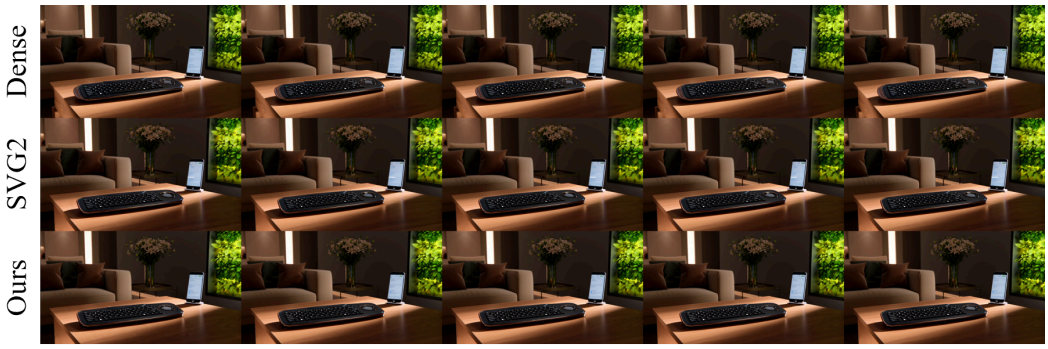


Figure 16: Additional qualitative comparisons of the dense attention, the Top-K baseline and our method.



Prompt: a keyboard and a cell phone.



Prompt: a kite and a baseball bat.



Prompt: A person is stretching leg.

Figure 17: Additional qualitative comparisons of the dense attention, SVG2 and our method.



Prompt: A person is drawing.



Prompt: a potted plant and a tv.



Prompt: A tranquil tableau of an exquisite mahogany dining table.

Figure 18: Additional qualitative comparisons of the dense attention, SVG2 and our method.