
Re-ttention: Ultra Sparse Visual Generation via Attention Statistical Reshape

Ruichen Chen
ECE Department
University of Alberta
ruichen1@ualberta.ca

Keith G. Mills
Division of CSE
Louisiana State University
keith.mills@lsu.edu

Liyao Jiang
ECE Department
University of Alberta
liyao1@ualberta.ca

Chao Gao
Huawei Technologies
Edmonton, Alberta, Canada
chao.gao4@huawei.com

Di Niu
ECE Department
University of Alberta
dniu@ualberta.ca

Abstract

Diffusion Transformers (DiT) have become the de-facto model for generating high-quality visual content like videos and images. A huge bottleneck is the attention mechanism where complexity scales quadratically with resolution and video length. One logical way to lessen this burden is sparse attention, where only a subset of tokens or patches are included in the calculation. However, existing techniques fail to preserve visual quality at extremely high sparsity levels and might even incur non-negligible compute overheads. To address this concern, we propose Re-ttention, which implements very high sparse attention for visual generation models by leveraging the temporal redundancy of Diffusion Models to overcome the probabilistic normalization shift within the attention mechanism. Specifically, Re-ttention reshapes attention scores based on the prior softmax distribution history in order to preserve the visual quality of the full quadratic attention at very high sparsity levels. Experimental results on T2V/T2I models such as CogVideoX and the PixArt DiTs demonstrate that Re-ttention requires as few as 3.1% of the tokens during inference, outperforming contemporary methods like FastDiTAttn, Sparse VideoGen and MInference.

1 Introduction

Diffusion Transformers (DiT) [33, 3, 2, 23, 10] combine the attention [38] mechanism with the iterative denoising of Diffusion Models [35] to generate high-quality visual content such as videos [47, 54] and images [20, 44, 41, 31]. However, a key bottleneck to generating longer videos and higher resolution content is the global properties of the self-attention module, whose compute cost scales quadratically with sequence size, i.e., resolution and video length.

Sparse attention techniques [6] aim to lower the computational burden by reducing the number of sequence tokens/patches that the attention mechanism considers during inference. Contemporary techniques like MInference [18] and Sparge Attention [51], as well as XAttention [46] achieve $\sim 50\%$ sparsity (i.e., reducing only 50% of the attention computations) by relying on downsampling the attention map or anti-diagonal scoring, respectively. In parallel, several recent methods [42, 48, 50] have been proposed specifically for DiTs, increasing the attention sparsity to $\sim 70\%$ by exploiting the inherent characteristics of diffusion process such as the progressively denoising structure and the spatial/temporal locality of attention.



Figure 1: Visual comparison using CogVideoX-2B [47] T2V model. Columns correspond to different frames. Rows correspond to different sparse attention methods (sparsity degree in paranthesis; higher is better). Prompt: “a colorful butterfly perching on a bud”. More examples in the Appendix.

While these methods can reduce more than half of the attention computation, their effectiveness remains limited for the growing computational demands of high-resolution image and video generation. Previous researches like LongFormer [1] and BigBird [49] can achieve >95% sparse attention. However, their reliance on retraining and fine-tuning introduces significant computation burdens, limiting their applicability to modern large-scale, pretrained generative models. Thus, the development of sparse attention techniques that achieve >95% sparsity with minimal visual quality loss remains an open challenge.

In this paper, we propose an effective method to statistically reshape the attention distribution distorted by the deployment of sparse attention, which we call Re-ttention. Re-ttention overcomes the high sparsity challenge faced by the training-free sparse attention method. Moreover, it is simple to implement and incurs negligible overhead compared to standard sparse attention at the same sparsity level. Figure 1 provides sample content from our technique compared to other sparse attention methods. Our detailed contributions are as follows:

1. We relate the failure to achieve degradation-free >95% sparsity without training from scratch to the distributional shift in attention scores caused by the reduced softmax denominator term, i.e., the row-wise sum of the exponentials of involved elements. We design an experiment to illustrate the importance of preserving this term and the impact on visual generation.
2. We discover the softmax distribution redundancy among neighboring denoising steps. Although the actual value of denominator changes unpredictably, the ratio between the sparse and full denominator is relatively stable.
3. We propose that the attention scores shifted by sparse attention are viable to be recovered by approximating the *real* softmax denominator.
4. The recovered attention scores deviate from a valid probability distribution, as their sum is less than one, violating the normalization property of softmax. We leverage the redundancy among neighboring denoising steps to compensate the missing probability with residual.

We apply Re-ttention to T2V models such as CogVideoX [47] in order to outperform contemporary methods like FastDiTAttn [48], Sparse VideoGen [42] and MInference [18] on relevant tasks such as VBench [16]. Furthermore, we apply Re-ttention to T2I models like the PixArt series [3, 2] and others [23] to maximize performance on Human Preference Score v2 (HPSv2) [41] and GenEval [12] while achieving a high sparsity of 96.9%.

2 Related Work

Diffusion Models (DM) [14] dominate visual generation tasks. Early DMs [35, 34] use convolutional U-Net structures [36] as their backbones. Later, Diffusion Transformers (DiT) [33, 5] adopt the attention-based [38] of Vision Transformers (ViT) [9] to increase scalability and visual generation quality. In addition to being the favored backbone structure for text-to-image (T2I) DMs [3, 2, 23, 10, 20, 44], the DiT structure is extensible to video generation [40, 19] as well. Specifically, Latte [30] proposes a 2D+1D attention block for video generation, which performs spatial and temporal attention separately. Subsequent works like CogVideoX [47] and OpenSora [24] adopt a 3D attention structure which processes the spatial and temporal dimensions simultaneously, yielding improved generation quality. However, this enhancement comes at the cost of significantly increased computation due to the quadratic complexity of attention, highlighting the pressing need for more efficient and sparse attention, which we explore in this work.

Sparse Attention denotes a class of techniques that aim to alleviate the hardware cost of the attention mechanism by omitting computation for unnecessary query-key pairs. Specifically, it is well documented that the attention mechanism produces sparse results [8, 28], yet suffers from a burdensome quadratic complexity and wasted computation by default. LongFormer [1] proposes sliding window attention that restricts attention to a local region. BigBird [49] and Mistral-7B [17] extend this idea to fine-grained attention masks, while SwinFormer [29] use local attention for efficient ViT design. Although these methods can reduce the attention computation by a factor of $8\times$ or more, they often necessitate training or fine-tuning the model, thus restricting the scope of deployment.

There are also training-free sparse attention methods. MInference [18] downsamples the attention probability matrix (QK^T) into blocks then dynamically select the top- k blocks to perform sparse attention. Subsequent research like FlexPrefill [21], Sparge Attention [51] and XAttention [46] rely on the block selection idea and propose dynamic block sorting algorithms. Further methods like StreamingLLM [43], DiTFastAttn [48] and Sparse VideoGen [42] identify the special attention patterns in LLM and DiT and propose efficient attention masking based on those patterns. However, the sparsity achievable by these methods is limited to $< 70\%$, which is much higher compared to prior works that require re-training or fine-tuning. We aim to address this gap and provide a training-free sparse attention method that can achieve $> 95\%$ sparsity on visual generation tasks.

Caching is a technique used in computer systems to temporarily store data or computations, thereby reducing redundant processing and improving overall efficiency. In DiT, the lengthy denoising process makes it well-suited for the application of caching techniques. Recent methods [4, 27, 53] re-use the attention outputs or the intermediate features at different denoising timesteps to skip the attention computation. Methods like DiTFastAttn [48, 50] leverage the caching mechanism to improve the visual quality.

3 Background: The Attention Mechanism and Sparsity

The attention mechanism [38] is the foundation of transformer architectures like DiTs. Let $X \in \mathbb{R}^{T \times d}$ be an input token/patch sequence, where T is the sequence length, dependent on the input size (e.g., image resolution or video length) and d is the embedding dimension, a hyperparameter of the transformer model. The attention mechanism contains h heads such that $d_h = \frac{d}{h}$; $d_h \in \mathbb{Z}^+$. We first map X into three representations, Query (Q), Key (K) and Value (V) of identical size $\mathbb{R}^{h \times T \times d_h}$, then compute the attention as follows,

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right)V. \quad (1)$$

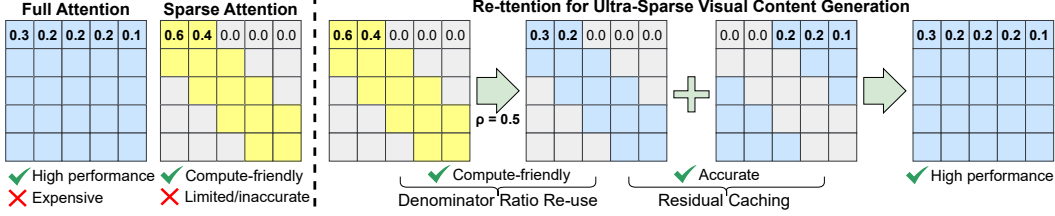


Figure 2: Illusion of attention map A computed by full attention, contemporary sparse attention (window-based) and our proposed Re-ttention. Sparse attention shifts the distribution of attention scores, resulting in degraded performance as sparsity increases. In contrast, Re-ttention re-uses the denominator ratio cached from the previous denoising steps to scale the sparse attention score to the full attention level. Then, we apply residual caching to accurately restore the full attention scores.

We can decompose this mechanism into several intermittent matrices, specifically the product of Q and K before (A^{pre}) and after (A) the softmax operation:

$$A^{\text{pre}} = \frac{QK^T}{\sqrt{d_h}} \in \mathbb{R}^{h \times T \times T}, \quad (2) \quad A = \text{Softmax}(A^{\text{pre}}) \in [0, 1]^{h \times T \times T}. \quad (3)$$

The computation of these matrices is very expensive [11]. To make matter worse, their size scales quadratically with T , which depends on the image/video resolution and video length. However, the softmax operation is computed row-wise and produces a probability distribution $\sum_{j=1}^T A_{:,j,j} = 1$, which empirically produces a sparse A in practice [8, 28].

Therefore, one way to alleviate this computational burden is to use a sparse attention mechanism. The key idea is to *omit* less relevant values of A^{pre} , that are likely to be 0 or close to 0 in A , from the softmax computation altogether. Formally, we express the sparse attention calculation using a mask $M \in \{0, 1\}^{h \times T \times T}$ where 1 means an index of A^{pre} will be included in the softmax, while the rest are excluded. The indexes of the included values in A^{pre} form a set $\mathcal{S} = \bigcup_{k=1}^h \bigcup_{i=1}^T \mathcal{S}_{k,i}$, where $\mathcal{S}_{k,i} = \{(k, i, j) | M_{k,i,j} = 1, 0 \leq j \leq T\}$.

Given an arbitrary element of the pre-softmax matrix $A_{k,i,j}^{\text{pre}}$, the normal and sparse softmax computation are given by

$$A_{k,i,j} = \frac{\exp(A_{k,i,j}^{\text{pre}})}{\sum_{t=1}^T \exp(A_{k,i,t}^{\text{pre}})}, \quad (4) \quad A_{k,i,j} = \begin{cases} \frac{\exp(A_{k,i,j}^{\text{pre}})}{\sum_{t \in \mathcal{S}_{k,i}} \exp(A_{k,i,t}^{\text{pre}})} & \text{if } j \in \mathcal{S}_{k,i}, \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

respectively. Ultimately, the mask matrix M determines the potential amount of computational savings. M can be computed statically [7] prior to inference or dynamically [18, 21, 51, 46] at runtime. Static techniques make more assumptions about the sparse regions of A while dynamic techniques impose additional inference overhead to compute M .

Regardless of technique, we can quantify the attention sparsity as a percentage, e.g., 10%, 50%, 90%, etc., simply by computing the ratio of values in M that are 0 as follows:

$$\text{Sparsity} = (1 - \frac{|\mathcal{S}|}{hT^2}) \times 100\%, \quad (6)$$

where a higher value for sparsity corresponds to a lower computational burden. Therefore, sparse M corresponds to an overall sparse attention. However, high sparsification can cause significant shifts in the softmax calculation statistics [43] and lead to detrimental performance. As we will next show, our proposed method, Re-ttention, aims to identify these statistical issues and address them.

4 Proposed Method: Re-ttention

In this section we form a hypothesis regarding how distributional shift in softmax statistics prevents current training-free sparse attention methods from satisfactorily operating at high sparsity, e.g.,

$> 95\%$. We then elaborate on our proposed Re-ttention technique, which overcomes this burden by re-using and caching softmax statistics at high sparsity. Figure 2 provides a high-level overview of our proposed technique in comparison to full and sparse attention.

4.1 Importance of the Softmax Denominator

As a preliminary investigation, we gauge the performance of several existing sparse attention techniques [18, 48, 42]. We consider the GenEval [12] benchmark and evaluate performance across a spectrum of sparsity values, i.e., starting at the highest sparsity these techniques consider in their original manuscript and then further increasing the sparsity.

Figure 3 illustrates our findings. We observe that existing approaches suffer a monotonic performance drop when the sparsity is further increased beyond their proposed value (denoted with ★). Per Eq. 5, a higher value of sparsity corresponds the inclusion of fewer tokens in the softmax denominator as $S_{k,i}$ shrinks. Thus, the further removal of tokens, i.e., increasing sparsity closer to 100%, has a larger impact on the overall denominator value [43]. This phenomenon, introduces a detrimental distribution shift in the overall attention scores.

We design a toy experiment to test this hypothesis and showcase the significance of the softmax denominator term. Specifically, we define a post-softmax masking operation as

$$A' = A \circ M, \quad (7)$$

where A is the output of the original full softmax attention via Eq. 4, $\circ$ denotes element-wise multiplication and A' is masked attention. We emphasize that Equation 7 is not a proper sparse attention calculation and does not entail speedup. *However*, it *mimics* the output of sparse attention as M still zeroes out the same indices of A , yet preserves the denominator of the full softmax.

We calculate M using sliding window attention [1]. We then generate visual content using both the formal sparse attention from Equation 5 and our post-softmax Equation 7 for comparison. Figure 4 provides a comparison, though we provide additional examples and prompts in the supplementary due to space constraints. We observe how the post-softmax attention preserves the guitar-playing panda, chair and background while the pre-softmax attention creates a noisy frame with jumbled contents where the panda appears to eat the guitar. Thus, these visual results validate our assumption regarding the importance of maintaining the softmax denominator. The challenge now becomes how to preserve this information in an efficient sparse attention setup.

4.2 Leveraging Denoising Properties for Statistical Reshape

One way to mitigate the distribution shift is to maintain the softmax denominator from the full attention calculation. We achieve this by exploiting the sequential nature of the DM denoising process and taking inspiration from DiT caching [4, 27, 53] and redundancy [37] methods.

Denominator Approximation. Figure 5 tracks the magnitude of the softmax denominator for a single token in the 9th head of the 12th DiT block in PixArt- α . Specifically, we calculate the

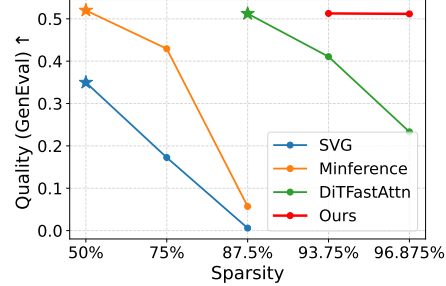


Figure 3: Quality-sparsity comparison of Re-ttention, Sparse VideoGen (SVG), MInference and DiTFastAttn. ★ denotes the sparsity level that prior methods operate under non-degraded conditions.



Figure 4: Visual comparison of pre-Softmax and post-Softmax masking on CogVideoX-2B with 66% sparsity, using sliding-window attention [1].

denominator using both the full and sparse attention (with 87.5% sparsity) as well as the ratio ρ between these statistics,

$$\rho = \frac{\sum_{t \in S_{k,i}} \exp(A_{k,i,t}^{\text{pre}})}{\sum_{t=1}^T \exp(A_{k,i,t}^{\text{pre}})}. \quad (8)$$

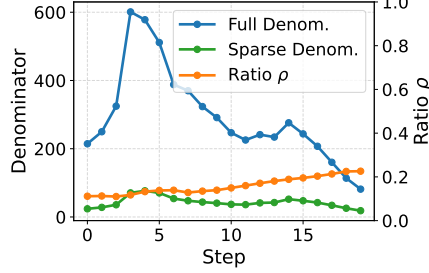


Figure 5: Plotting softmax denominators for full and sparse attention as well as the ratio ρ per Eq. 8 across 20 steps.

This yields an insightful observation: While the actual value of the denominators may change unpredictably and non-monotonically over the denoising process, given a statically computed mask M , the ratio ρ follows a predictable trend. Therefore, we propose a simple method to recover the attention distribution by modifying the output of the sparse softmax operation Eq. 5 as follows:

$$A'_{k,i,j} = \rho A_{k,i,j}, \quad (9)$$

where we cache $0 < \rho \leq 1$ from a previous step. Multiplying the Softmax output of sparse attention by ρ approximates the full attention value, mitigating the distribution shift. In practice, since ρ is not a constant per Fig. 5, we empirically find that slightly increasing the cached ρ after each denoising step achieves better performance. Thus, we parameterize $\rho_t = \rho_{t-1} + \lambda$ after each denoising step, where λ is a ramp-up hyperparameter.

Residual Caching. Although we can modify the softmax to make the sparse attention more closely match that of full attention, Equation 8 does not yield proper probability distributions as $\sum_{j=1}^T A_{:,j} < 1$. Practically, this reduces the magnitude of overall attention outputs per Equation 1, which can negatively impact performance. To address this issue, we first define the residual R as the difference between the full attention computed via Equation 4 and our ρ -reshaped attention via Equations 5 and 9:

$$R = \text{FullAttention}(Q, K, V) - \text{ReshapeAttention}(Q, K, V, \rho). \quad (10)$$

We can later add R to our sparse attention output. In fact, R is mathematically equivalent to the attention output of the masked tokens at the caching timestep.

Overall, the idea behind Re-ttention is to compute sparse attention for important regions, while re-using previously cached statistics from previous steps in less important regions of the attention map. This involves caching necessary statistics from the full attention in some steps, which is common in DiT sparse attention methods [48, 50].

5 Experimental Setup and Results

We evaluate Re-ttention on both the text-to-video (T2V) and text-to-image (T2I) tasks using a number of DiT models, such as CogVideoX (2B) [47], PixArt- α/Σ (0.6B) [3, 2] and Hunyuan-DiT (1.6B) [23]. We generate 720×480 resolution, 6 second videos at 8 fps and 1024×1024 pixel images throughout this paper. We compare to several existing sparse attention methods for visual content in the literature like Sparse VideoGen (SVG) [42], MInference [18] and DiTFastAttn [48, 50] to demonstrate both qualitative and quantitative performance gains and computational cost savings.

Implementation Details. Specifically, we use the HuggingFace Diffusers library [39] to instantiate the base DiT models and consider the default values for inference parameters like the classifier-free guidance (CFG) scale and number of denoising steps - 50 for CogVideoX/Hunyuan and 20 for the PixArt DiTs. Following prior literature on DiT acceleration [42, 53, 22, 27], we apply the full attention during the first 5, 10 or 15 steps for the PixArt DiTs, Hunyuan and CogVideoX models, respectively, and then apply the sparse attention mechanism for the remainder of the denoising process. Further, we set a caching period of 5 steps for DiTFastAttn and Re-ttention, where we

Table 1: Quantitative evaluation results for T2V model CogVideoX-2B [47] on VBench [16] and other metrics. Arrows indicate if a higher or lower value of a metric is preferred. Best and second-best results in **bold** and *italics*, respectively.

Attention	Sparsity $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	ImageQual $\uparrow$	SubConsist $\uparrow$
Full-Attention	0.0%	Reference	Reference	Reference	65.72%	94.97%
SVG	87.5%	14.48	0.548	0.501	54.48%	89.26%
SVG	96.9%	10.50	0.418	0.898	51.82%	96.73%
MInference	87.5%	14.99	0.558	0.480	53.78%	83.71%
MInference	96.9%	9.25	0.325	0.818	34.36%	75.84%
DiTFastAttn	96.9%	27.93	<i>0.865</i>	<i>0.098</i>	<i>64.86%</i>	94.32%
Re-ttention	96.9%	27.96	0.894	0.059	64.87%	94.80%

perform full attention to cache required statistics. For fair comparison, we apply this caching to SVG and MInference as well: SVG and MInference will perform full attention at the same timesteps as DiTFastAttn and Re-ttention. To perform T2I using SVG, we treat the image as a video containing a single frame. We provide further baseline experimental details in the supplementary.

The rest of this section is organized as follows: We enumerate our T2V and T2I evaluation setup and results in Sections 5.1 and 5.2, respectively. Next, we provide ablation studies in Section 5.3.

5.1 Text-to-Video Evaluation

We perform quantitative T2V evaluation using the Animal and Architecture categories of VBench [16], which consist of 100 videos each. For video quality, we use VBench score to evaluate standalone video quality. Specifically, we follow previous literature [42] and report the Image Quality and Subject Consistency metrics in VBench. Additionally, we compute the Peak Signal-to-Noise Ratio (PSNR) [15], Structural Similarity Index Measure (SSIM) [32] and Learned Perceptual Image Patch Similarity (LPIPS) [52]. These metrics evaluate the similarity and quality of videos generated by sparse attention methods relative to those generated using the full attention mechanism. We evaluate all methods at 96.9% sparsity to provide an apples-to-apples performance investigation. However, some methods exhibit substantial degradation at this level and produce very noisy/black frames, so we additionally report results at a less aggressive setting of 87.5% sparsity.

Table 1 presents our findings. Results demonstrate that Re-ttention consistently outperforms all other baselines in terms of video quality and similarity metrics. Notably, Re-ttention not only outperforms both SVG and MInference at the strict sparsity of 96.9%, but also at 87.5% sparsity. The one exception is SVG at 96.9% sparsity, which achieves the highest SubConsist performance. However, this result is an outlier, as it even exceeds the SubConsist performance of full attention significantly while SVG substantially underperforms on all other metrics at this sparsity level. Furthermore, Re-ttention also outperforms DiTFastAttn, which also involves caching additional statistics at the high sparsity level of 96.9%. Therefore, overall, these results demonstrate the robustness, competitive performance of Re-ttention at $> 95\%$ sparsity in T2V applications.

We provide some sample frames from videos generated by Re-ttention, baseline sparse attention methods and full attention. Specifically, recall Figure 1 in the introduction. The video generated by Re-ttention shows the best clarity and temporal consistency across frames for the main subject, and it has no artifacts in the background. Moreover, the video generated by Re-ttention is most similar to the reference video generated with full-attention. In contrast, the video generated by DiTFastAttn has noisy texture artifacts both in the background and the subject. For SVG and MInference, the subject is inconsistent and deformed despite using a much lower sparsity. We provide more T2V visual comparisons in the supplementary materials.

5.2 Text-to-Image Results

We evaluate T2I performance on a comprehensive set benchmark metrics: GenEval [12], HPSv2 [41], and MS-COCO 2014 [25]. GenEval consists of 553 unique prompts. For each prompt, the DiT generates 4 images. HPSv2 consists of four image categories: Animation, Concept-art, Painting and Photos. Each category consists of 800 images for 3.2k generations in total. Finally, we generate 10k

Table 2: Quantitative evaluation results for PixArt- α [3], PixArt- Σ [2] and Hunyuan-DiT [23] across the GenEval [12], HPSv2 [41], and MS-COCO 2014 [25] benchmarks. Best and second best results in **bold** and *italics*, respectively.

Model	Attention	Sparsity $\uparrow$	GenEval $\uparrow$	HPSv2 $\uparrow$	LPIPS $\downarrow$	IR $\uparrow$	CLIP $\uparrow$
PixArt- α	Full-Attention	0.0%	0.480	30.79	Reference	0.864	31.28
	SVG	75.0%	0.368	25.24	0.655	-0.141	29.43
	MInference	75.0%	0.433	<i>28.04</i>	0.458	0.549	30.93
	DiTFastAttn	93.8%	0.431	27.26	0.506	0.688	30.72
	DiTFastAttn	96.9%	0.364	26.71	0.590	0.314	29.63
	Re-ttention	93.8%	0.456	28.29	0.354	0.688	31.21
	Re-ttention	96.9%	<i>0.448</i>	<i>27.57</i>	<i>0.372</i>	<i>0.646</i>	<i>31.20</i>
PixArt- Σ	Full-Attention	0.0%	0.544	30.70	Reference	0.953	31.54
	SVG	75.0%	0.172	18.48	0.742	-1.315	26.09
	MInference	75.0%	0.429	27.09	0.536	0.457	30.76
	DiTFastAttn	93.8%	0.411	27.64	0.591	0.507	30.08
	DiTFastAttn	96.9%	0.233	22.79	0.734	-0.600	27.37
	Re-ttention	93.8%	0.513	28.37	0.417	0.808	31.59
	Re-ttention	96.9%	<i>0.512</i>	<i>27.72</i>	<i>0.435</i>	<i>0.784</i>	31.59
Hunyuan	Full-Attention	0.0%	0.610	30.41	Reference	1.027	31.77
	SVG	75.0%	0.317	24.73	0.854	-0.574	27.92
	MInference	75.0%	0.450	23.94	0.720	-0.063	30.15
	DiTFastAttn	93.8%	0.024	14.77	0.896	-2.074	22.47
	DiTFastAttn	96.9%	0.002	12.28	0.923	-2.237	22.10
	Re-ttention	93.8%	<i>0.585</i>	29.03	0.598	<i>0.911</i>	<i>31.63</i>
	Re-ttention	96.9%	0.590	28.89	<i>0.606</i>	0.923	31.65



Figure 6: Visual comparison on MS-COCO 2014 [25] prompts using PixArt- α (row 1), PixArt- Σ (row 2), and Hunyuan (row 3). We show images generated by Re-ttention (our method) and by other attention methods in different columns. We provide further examples in the appendix.

images using the MS-COCO 2014 validation set and measure the LPIPS score [52], ImageReward (IR) [45] and CLIP score [13] using the ViT-B/16 backbone.

Table 2 lists our results on the T2I task. Re-ttention outperforms all other sparse attention methods across models and metrics, showing consistently better performance. Additionally, Re-ttention achieves this while operating under an extremely high sparsity of 96.9%, which reduces the token/patch sequence to less than one *twentieth* of its original size, whereas other baseline methods underperform at 75% sparsity, which only reduces sequence length down to one *fourth*. Additionally, our performance on the IR metric is consistently positive, a feat that no other sparse attention method attains. Moreover, while DiTFastAttn attains similar T2V performance to Re-ttention, it fails to generalize to the T2I task on Hunyuan in terms of GenEval and HPSv2 performance, even at reduced sparsity of 93.8%. In contrast, Re-ttention performance neither suffers at 93.8% nor 96.9% sparsity, underscoring the effectiveness of our technique.

Next, we present the visual (qualitative) comparisons on PixArt- α [3], PixArt- Σ [2] and Hunyuan [23] T2I models in Figures 6. More visual comparisons can be found in the Appendix. Overall, Re-ttention generates images with better image quality than other sparse attention methods and has higher similarity to the reference images generated by full-attention, even when using an extreme sparsity of 96.9%. For PixArt- α [3] and PixArt- Σ [2], Re-ttention generates clean, high-quality images that are well aligned to the prompts. Whereas the other methods often generate colored noise artifacts, distorted subjects, and lower quality images. For Hunyuan [23], we observe that the other sparse attention methods generate severely degraded images, while Re-ttention can generate images that are similar to images generated by full-attention.

5.3 Ablation Studies

Finally, we ablate the effect of the ramp-up hyperparameter λ on PixArt- Σ [2] on overall performance. Specifically, we evaluate on HPSv2 overall as well as the ‘animation’ and ‘concept-art’ categories. Table 3 reports our findings. These findings demonstrate the robustness of Re-ttention as it is possible to forgo the λ parameter, yet it is better to select a moderate value.

Table 3: HPVs2 score under different ramp-up hyperparameter λ with 96.9% sparsity on PixArt- Σ [2].

λ	Anime $\uparrow$	ConceptArt $\uparrow$	HPSv2 $\uparrow$
0	29.40	26.94	27.46
0.01	28.89	26.21	26.83
0.02	28.88	26.19	26.82
0.04	29.60	27.23	27.72

6 Conclusions and Future Work

We propose Re-ttention, a training-free sparse attention method for Diffusion Transformers, which achieves 96.9% sparsity without performance loss on DiTs like CogVideoX and Hunyuan. We attain these gains by identifying the distribution shift of attention scores incurred by sparse attention methods that prevents extreme sparsity ($> 95\%$) without significant performance degradation and resolve this issue using a combination of caching and statistical re-use. We evaluate Re-ttention on T2V and T2I tasks, outperforming contemporary baselines like SVG, MInference and DiTFastAttn.

Potential future directions to expand Re-ttention and address limitations should aim to repurpose our contributions in the application domain of LLMs or autoregressive visual content generation models. These models rely on causally masked attention, meaning that our attention statistical reshape, which leverages the step-wise denoising process in diffusion models to reuse cached attention statistics from previous steps, must be handled differently in this setting where such sequential caching is unavailable. Also, Re-ttention implements sparse attention using a static mask, though further investigation is merited to validate it in the context of dynamically-generated sparse attention masks. Dynamically adapting the sparsity pattern based on attention statistics or token importance could further improve efficiency while preserving output quality, enabling Re-ttention to generalize across a wider range of sequence modeling and generative tasks.

References

- [1] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.
- [2] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- Σ : Weak-to-strong training of diffusion transformer for 4k text-to-image generation, 2024.
- [3] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=eAKmQP3m1>.
- [4] Pengtao Chen, Mingzhu Shen, Peng Ye, Jianjian Cao, Chongjun Tu, Christos-Savvas Bouganis, Yiren Zhao, and Tao Chen. Delta-dit: A training-free acceleration method tailored for diffusion transformers. *arXiv preprint arXiv:2406.01125*, 2024.
- [5] Ruichen Chen, Keith G Mills, and Di Niu. Fp4dit: Towards effective floating point quantization for diffusion transformers. *arXiv preprint arXiv:2503.15465*, 2025.
- [6] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.
- [7] Steve Dai, Hasan Genc, Rangharajan Venkatesan, and Brucek Khailany. Efficient transformer inference with statically structured sparse attention. In *2023 60th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2023.
- [8] Yichuan Deng, Zhao Song, and Chiwun Yang. Attention is naturally sparse with gaussian distributed input. *arXiv preprint arXiv:2404.02690*, 2024.
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.
- [11] Yao Fu. Challenges in deploying long-context transformers: A theoretical peak performance analysis. *arXiv preprint arXiv:2405.08944*, 2024.
- [12] Dhruva Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.
- [13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [15] Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369, 2010. doi: 10.1109/ICPR.2010.579.
- [16] Ziqi Huang, Yanan He, Jiahuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.
- [17] AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b. arxiv 2023. *arXiv preprint arXiv:2310.06825*, 2024.
- [18] Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir Abdi, Dongsheng Li, Chin-Yew Lin, et al. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. *Advances in Neural Information Processing Systems*, 37:52481–52515, 2024.

- [19] 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.
- [20] Black Forest Labs. flux, 2024. URL <https://github.com/black-forest-labs/flux>.
- [21] Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, and Xun Zhou. Flexprefill: A context-aware sparse attention mechanism for efficient long-sequence inference. *arXiv preprint arXiv:2502.20766*, 2025.
- [22] Muyang Li, Tianle Cai, Jiaxin Cao, Qinsheng Zhang, Han Cai, Junjie Bai, Yangqing Jia, Kai Li, and Song Han. Distrifusion: Distributed parallel inference for high-resolution diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7183–7193, 2024.
- [23] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xincheng Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024.
- [24] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.
- [25] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.
- [26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [27] Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It’s time to cache for video diffusion model. *arXiv preprint arXiv:2411.19108*, 2024.
- [28] Liu Liu, Zheng Qu, Zhaodong Chen, Fengbin Tu, Yufei Ding, and Yuan Xie. Dynamic sparse attention for scalable transformer acceleration. *IEEE Transactions on Computers*, 71(12):3165–3178, 2022.
- [29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [30] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.
- [31] Keith G. Mills, Mohammad Salameh, Ruichen Chen, Wei Hassanpour, Negar Lu, and Di Niu. Qua²sedimo: Quantifiable quantization sensitivity of diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6153–6163, 2025.
- [32] Jim Nilsson and Tomas Akenine-Möller. Understanding ssim. *arXiv preprint arXiv:2006.13846*, 2020.
- [33] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.
- [34] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=di52zR8xgf>.
- [35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

- [37] Xibo Sun, Jiarui Fang, Aoyu Li, and Jinzhe Pan. Unveiling redundancy in diffusion transformers (dits): A systematic study. *arXiv preprint arXiv:2411.13588*, 2024.
- [38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- [39] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>, 2022.
- [40] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwei Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.
- [41] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.
- [42] 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.
- [43] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.
- [44] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformer, 2024. URL <https://arxiv.org/abs/2410.10629>.
- [45] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 15903–15935, 2023.
- [46] Ruyi Xu, Guangxuan Xiao, Haofeng Huang, Junxian Guo, and Song Han. Xattention: Block sparse attention with antidiagonal scoring. *arXiv preprint arXiv:2503.16428*, 2025.
- [47] 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.
- [48] Zhihang Yuan, Hanling Zhang, Lu Pu, 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.
- [49] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.
- [50] Hanling Zhang, Rundong Su, Zhihang Yuan, Pengtao Chen, Mingzhu Shen Yibo Fan, Shengen Yan, Guohao Dai, and Yu Wang. Ditfastattnv2: Head-wise attention compression for multi-modality diffusion transformers. *arXiv preprint arXiv:2503.22796*, 2025.
- [51] Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. Sparsgeattn: Accurate sparse attention accelerating any model inference. *arXiv preprint arXiv:2502.18137*, 2025.
- [52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
- [53] Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. Real-time video generation with pyramid attention broadcast. *arXiv preprint arXiv:2408.12588*, 2024.
- [54] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist”,**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

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

Answer: [Yes]

Justification: The claims are stated in abstract and introduction. And they match the experimental results in the results section.

Guidelines:

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

2. Limitations

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

Answer: [Yes]

Justification: The limitations are discussed in the appendix due to space constraints.

Guidelines:

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

3. Theory assumptions and proofs

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

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

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

4. Experimental result reproducibility

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

Answer: [Yes]

Justification: They are described in detail in the results section.

Guidelines:

- The answer NA means that the paper does not include experiments.

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

5. Open access to data and code

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

Answer: [\[Yes\]](#)

Justification: Code is included as supplementary material.

Guidelines:

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

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

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

Answer: [\[Yes\]](#)

Justification: The experiment settings and implementation details are provided in the results section.

Guidelines:

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

7. Experiment statistical significance

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

Answer: [\[No\]](#)

Justification: This is not standard practice in the literature for the field of research this paper targets. Instead, we perform extensive evaluation on a range of models, and tasks and show different performance metrics.

Guidelines:

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

8. Experiments compute resources

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

Answer: [\[Yes\]](#)

Justification: They are provided in the results section and appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.

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

9. Code of ethics

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

Answer: [Yes]

Justification: It conforms to the code of ethics in every aspect.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

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

Answer: [Yes]

Justification: It is included in the Appendix.

Guidelines:

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

11. Safeguards

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

Answer: [NA]

Justification: We are not releasing any new data or model. Our method is applied to accelerate existing models on existing data.

Guidelines:

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

12. Licenses for existing assets

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

Answer: [\[Yes\]](#)

Justification: Credits are given to all assets used.

Guidelines:

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

13. New assets

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

Answer: [\[NA\]](#)

Justification: The paper does not release new assets.

Guidelines:

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

14. Crowdsourcing and research with human subjects

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

Answer: [NA]

Justification: The paper does not involve any crowdsourcing or human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

Answer: [NA]

Justification: The paper does not involve any crowdsourcing or human subjects. Therefore, no IRB approval or equivalent review was required or obtained.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: The method does not involve any LLM as a component.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Appendix

We provide additional information about our work. Sections A.1 and A.2 provide statements about the broader impacts of our work and limitations, respectively. Further, we provide additional details on our methodology in Sec. A.3 and baselines in Sec. A.4. Section A.5 provides elaborates on our pre vs. post-softmax example from Figure 4. Finally, Sections A.6 and A.7 provide additional T2V and T2I results, respectively.

A.1 Societal Impacts

Re-ttention improves the efficiency of image and video generation by enabling extremely sparse attention. This makes high-quality generative models more accessible and environmentally sustainable by reducing computational and energy demands. By lowering resource barriers, Re-ttention can benefit creators, educators, and researchers in low-resource settings. While any generative model carries a risk of misuse, Re-ttention does not introduce new risks beyond existing systems. Responsible deployment and continued dialogue on ethical use remain important.

A.2 Limitation

We design Re-ttention around achieving high sparsity for the non-autoregressive self-attention mechanism utilized by visual generation DiTs, rather than the autoregressive, causally-masked attention of LLMs which may more often feature different attention patterns such as columns [7, 43]. Additionally, Re-ttention exploits the sequential nature of DMs and is inspired by DiT caching techniques [4, 27, 53]. Our method may not be readily generalizable to autoregressive LLMs, though modifications and expansions into this field are a potential future work. Furthermore, Re-ttention is currently designed for statically computed attention masks, which offer speedup advantages. Extending the approach to support dynamically computed masks to support fine-grained sparse attention presents a promising direction for future work.

Although we did not implement a custom GPU kernel, we measured inference latency on typical GPUs and observed that Re-tention achieves comparable runtime to DiTFastAttn across all tested models. This demonstrates that our contributions do not impose significant computational overhead, confirming that Re-tention maintains both high sparsity and practical efficiency.

A.3 Explanation of Re-ttention

In Section 4 we claim that the residual R in Re-ttention is mathematically equivalent to the attention output of the masked tokens at the caching timestep. We now further elaborate on this claim:

Recall the definition of A in Eq. 3 and the set S that contains the included values (by sparse attention) in A . Hence, the A can be decomposed into two parts:

$$\begin{aligned} A &= A_{\in S} + A_{\notin S}, \\ A_{\in S} &= A \circ \mathbf{1}_{(k,i,j) \in S}, \\ A_{\notin S} &= A \circ \mathbf{1}_{(k,i,j) \notin S}, \end{aligned} \tag{11}$$

where $\mathbf{1}_{\in S}$ is the indicator matrix that is 1 where $(k, i, j) \in S$, and 0 elsewhere. Conversely, $\mathbf{1}_{\notin S}$ is 1 where $(k, i, j) \notin S$ and 0 elsewhere.

At the caching timestep, we have the ratio ρ between the denominator of full and sparse attention according to Eq. 8. Because we compute full attention in the caching step, the ratio ρ is not an approximation but an *accurate* value. Hence, we have:

$$\text{ReshapeAttention}(Q, K, V, \rho) = \rho A \cdot V = A_{\in S} \cdot V \tag{12}$$

Therefore, the residual R in Eq. 10 is:

$$R = A \cdot V - A_{\in S} \cdot V = A_{\notin S} \cdot V, \tag{13}$$

which is mathematically equivalent to the attention output of the masked tokens at the caching timestep.

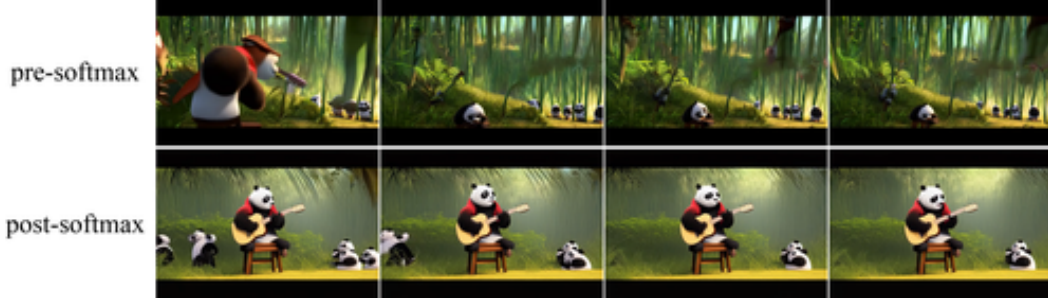


Figure 7: Visual comparison of pre-softmax and post-softmax masking on CogVideoX-2B with 66% sparsity. Prompt: “A panda, dressed in a small, red jacket and a tiny hat, sits on a wooden stool in a serene bamboo forest. The panda’s fluffy paws strum a miniature acoustic guitar, producing soft, melodic tunes. Nearby, a few other pandas gather, watching curiously and some clapping in rhythm. Sunlight filters through the tall bamboo, casting a gentle glow on the scene. The panda’s face is expressive, showing concentration and joy as it plays. The background includes a small, flowing stream and vibrant green foliage, enhancing the peaceful and magical atmosphere of this unique musical performance”.



Figure 8: Visual comparison of pre-softmax and post-softmax masking on CogVideoX-2B with 66% sparsity. Prompt: “A detailed wooden toy ship with intricately carved masts and sails is seen gliding smoothly over a plush, blue carpet that mimics the waves of the sea. The ship’s hull is painted a rich brown, with tiny windows. The carpet, soft and textured, provides a perfect backdrop, resembling an oceanic expanse. Surrounding the ship are various other toys and children’s items, hinting at a playful environment. The scene captures the innocence and imagination of childhood, with the toy ship’s journey symbolizing endless adventures in a whimsical, indoor setting”.

A.4 Details of Baseline Implementation

We compare Re-ttention to three different baseline methods: Sparse VideoGen (SVG) [42], MInference [18], and DiTFastAttn [48]. We enumerate the experiment implementation details for Text-to-Video (T2V) and Text-to-Image (T2I) generation tasks, respectively.

Text-to-Video For DiTs, MInference classifies all attention heads into a block sparse format [18] to generate M . We use the SVG official implementation for CogVideoX series to generate videos. As for Re-ttention and DiTFastAttn, we use sliding window attention, which restricts each token’s attention to a local neighborhood and will repeat the same mask at each frame of the video.

Text-to-Image Since T2I generation lacks a temporal dimension, we apply only the spatial attention heads in SVG and adjust the window size to match the target sparsity. Re-ttention and DiTFastAttn use the same sliding window attention as SVG.

A.5 Additional Examples for Post-Softmax Masking Operation

Figure 7 expands on Fig. 4 by providing additional video frame comparisons and the lengthy textual prompt. Post-Softmax masking not only better preserves the objects (panda, stool, guitar, etc.), but also consistently maintains the main part of the video over time, while pre-softmax causes the

large panda to vanish. This example further validates our assumption regarding the importance of maintaining the softmax denominator.

Further, Figure 8 provides an additional comparison with a different prompt. In the pre-softmax video, the ship becomes increasingly distorted over time, whereas in the post-softmax video, it remains consistent throughout. Notably, the reduced texture detail in the post-Softmax output reveals an issue caused by denormalized attention probabilities—specifically, information loss due to the sum of softmax probabilities being less than one, leading to a shrinkage effect in the features.

A.6 Visual Comparison for Video Generation

We show additional visual (qualitative) comparisons on video generation using the CogVideoX-2B [47] model in Figures 9, 10, 11 and 12. For example, in Figure 10, Re-ttention has the best looking otter as well as the food with the most similar shape as the reference video. Besides, while other baseline methods have artifacts like blurry textures and distortions in the background, Re-ttention preserves background fidelity, closely matching the reference video. Those additional comparisons match the experiment result in the main paper: The videos generated by Re-ttention are the most similar to the reference video generated by full-attention; also, it has the best clarity and consistency and no artifacts in the background.

A.7 Visual Comparison for Image Generation

We show additional visual (qualitative) comparisons on image generation using the PixArt- α [3], PixArt- Σ [2], and Hunyuan [23] models in Figures 13, 14 and 15, respectively. The main object generated by the dynamic sparse attention baseline MInference deviates significantly from the full-attention reference, often resulting in unnatural or distorted appearances. For static baseline methods like SVG and DiTFastAttn, although the main objects in their images are more similar to the full-attention reference images, there are artifacts in the background which degrade the image quality. In comparison, Re-ttention not only preserves the fidelity of the main object but also mitigates background artifacts, demonstrating superior performance in T2I generation and strong generalization across different DiT architectures.

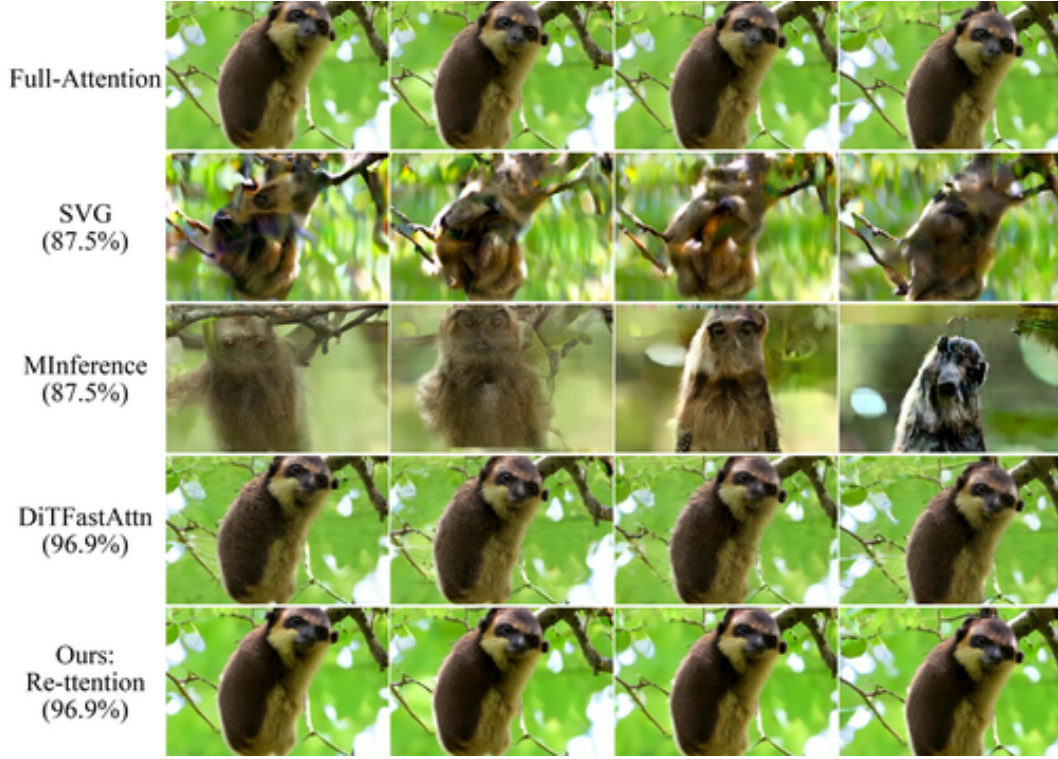


Figure 9: **T2V** visual comparison using CogVideoX-2B [47] T2V model. Each row corresponds to video frames generated by different methods. Prompt: “a curious sloth hanging from a tree branch”.

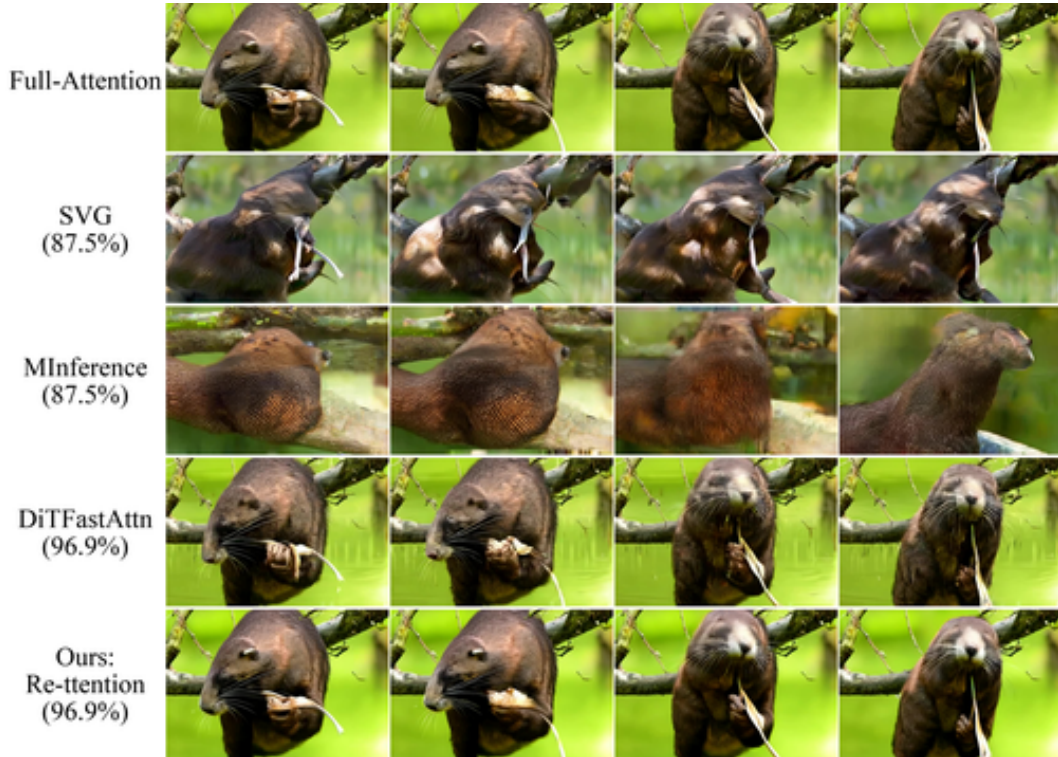


Figure 10: **T2V** visual comparison using CogVideoX-2B [47] T2V model. Each row corresponds to video frames generated by different methods. Prompt: “otter on branch while eating”.

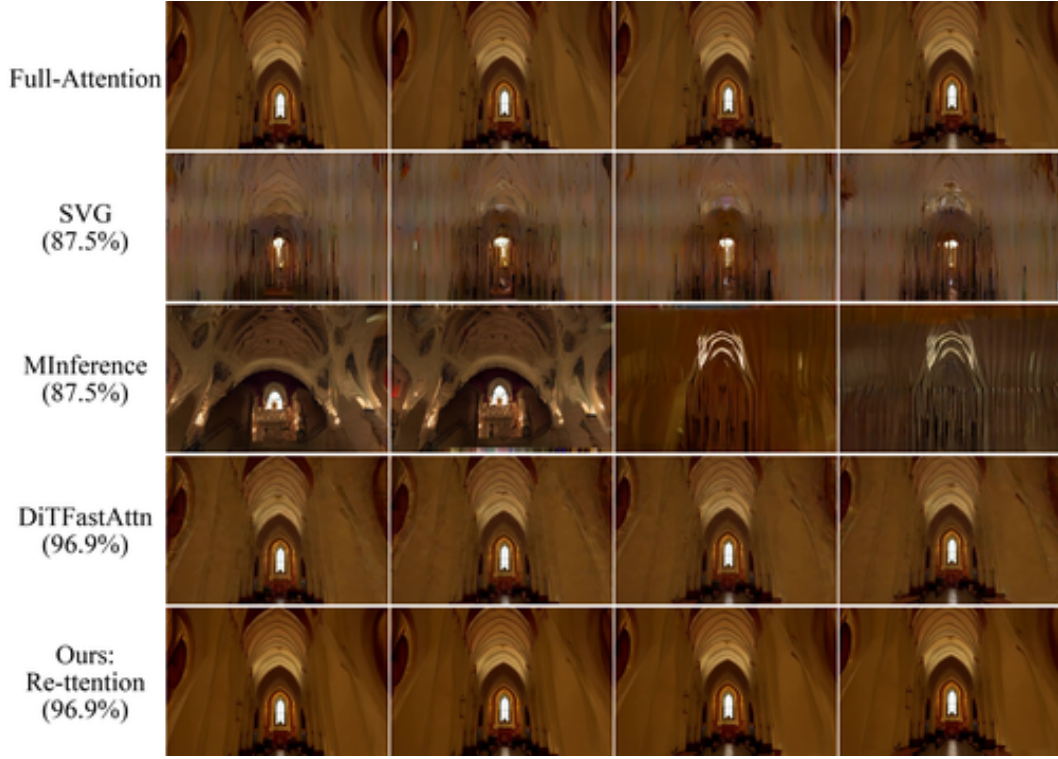


Figure 11: **T2V** visual comparison using CogVideoX-2B [47] T2V model. Each row corresponds to video frames generated by different methods. Prompt: “a church interior”.

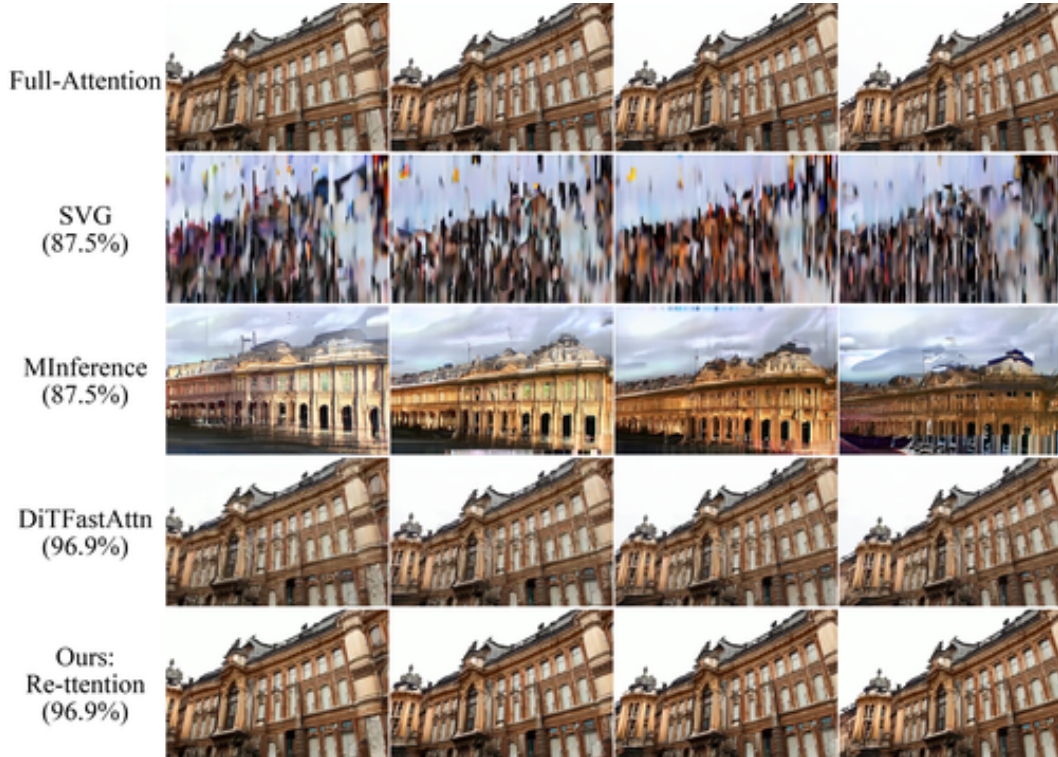


Figure 12: **T2V** visual comparison using CogVideoX-2B [47] T2V model. Each row corresponds to video frames generated by different methods. Prompt: “the georgian building”.

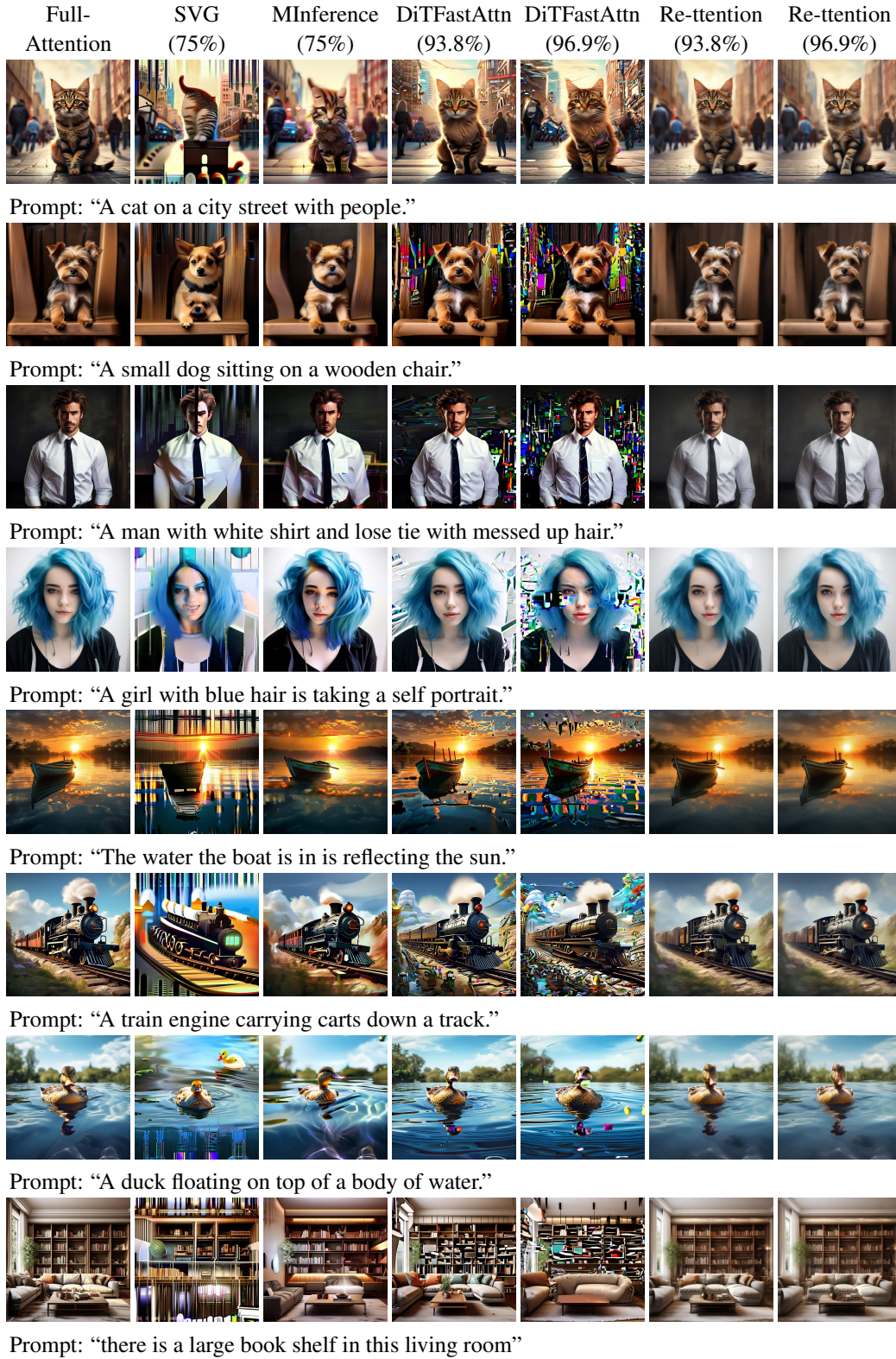


Figure 13: **T2I** visual comparison on MS-COCO 2014 [26] dataset using PixArt- α [3] model. Each row corresponds to one prompt, we show images generated by Re-ttention (our method) and by other attention methods in different columns.

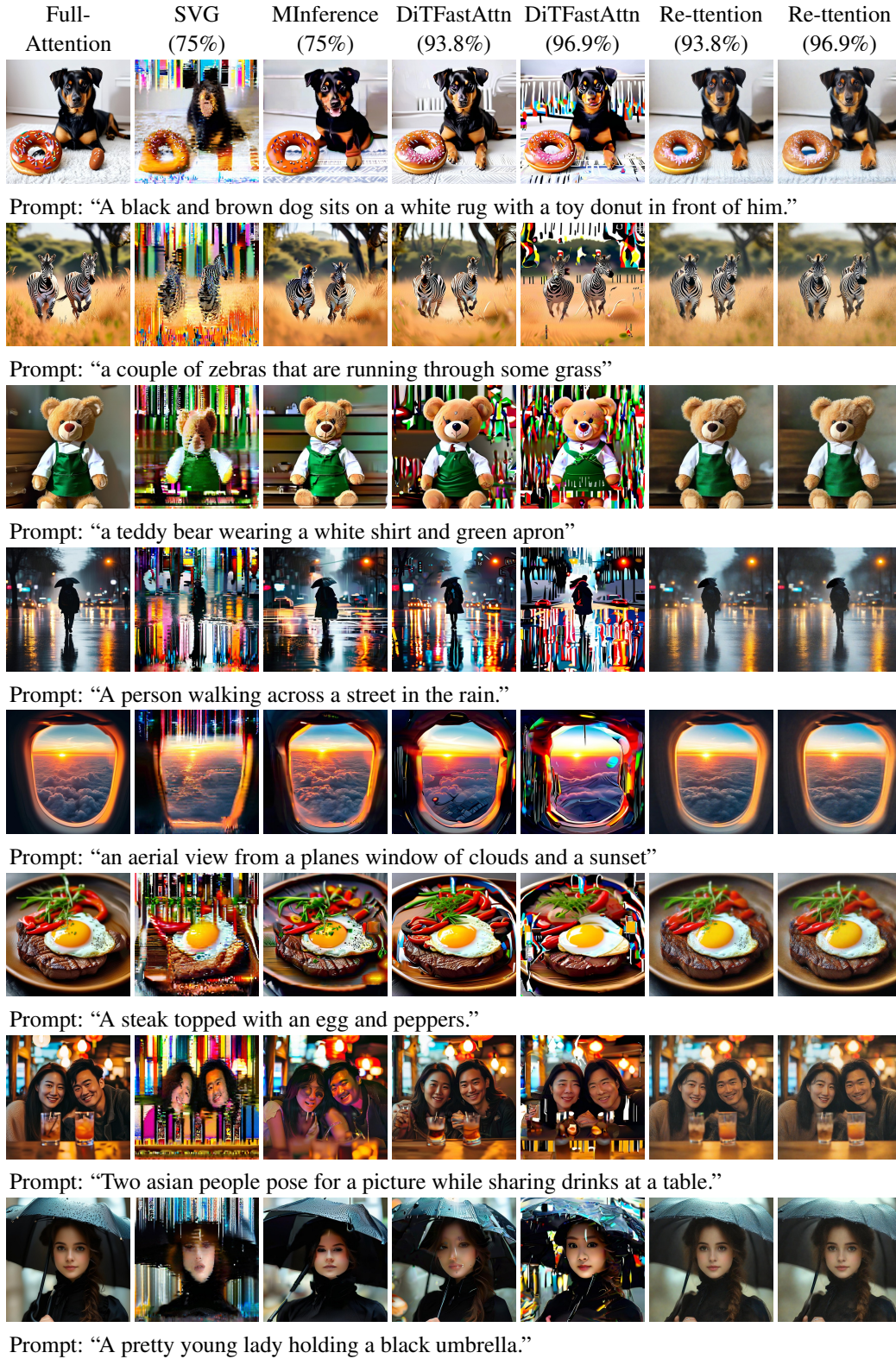


Figure 14: **T2I** visual comparison on MS-COCO 2014 [26] dataset using PixArt- Σ [2] model. Each row corresponds to one prompt, we show images generated by Re-ttention (our method) and by other attention methods in different columns.

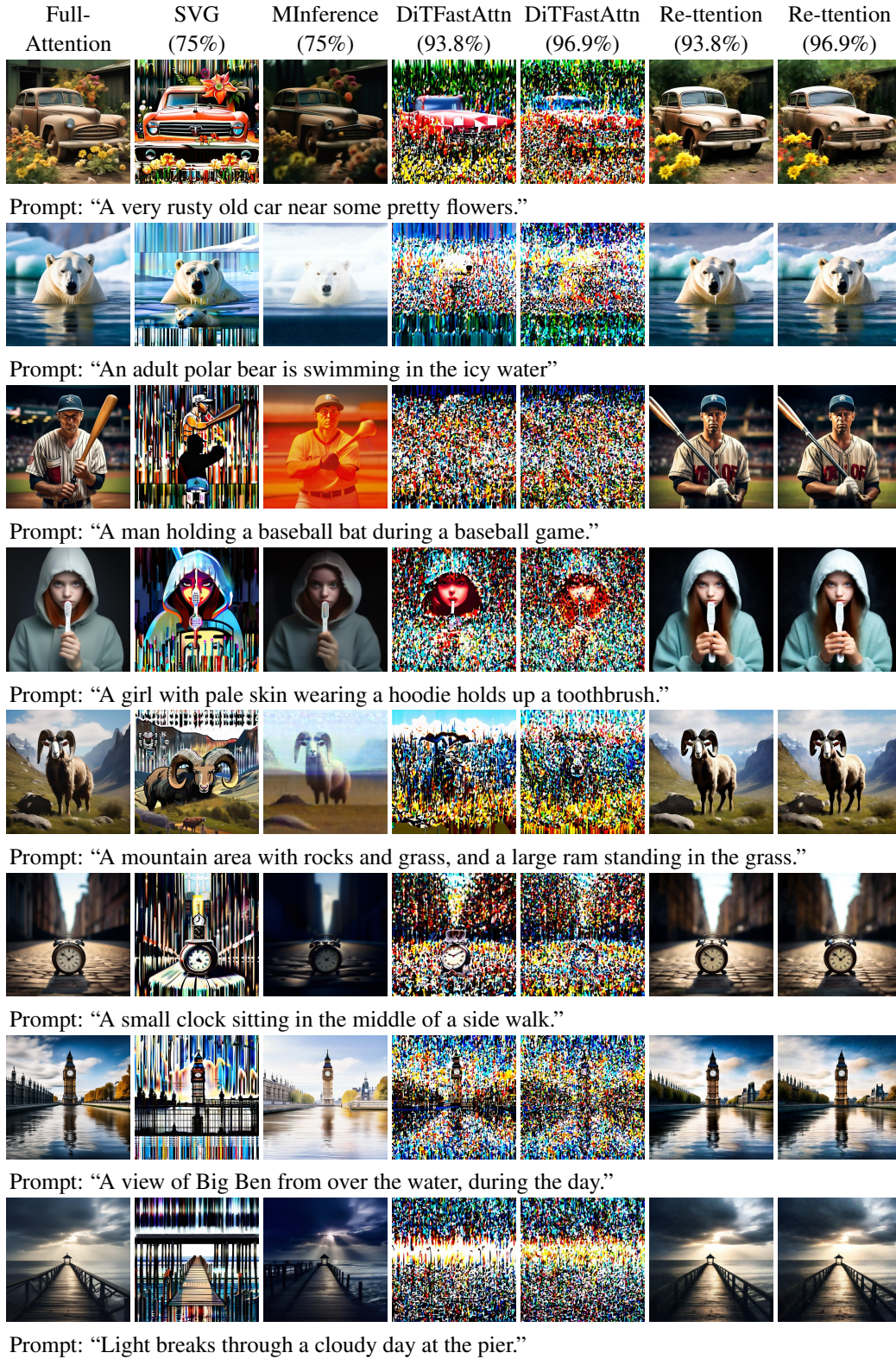


Figure 15: **T2I** visual comparison on MS-COCO 2014 [26] dataset using Hunyuan-DiT [23] model. Each row corresponds to one prompt, we show images generated by Re-ttention (our method) and by other attention methods in different columns.