Appendix

Comparison of MMDiT and DiT structure

Unlike DiT, as shown in Fig 8, MMDiT eliminates the cross-attention module in DiT; instead, visual inputs and text embeddings are independently projected and then concatenated together for self-attention processing.

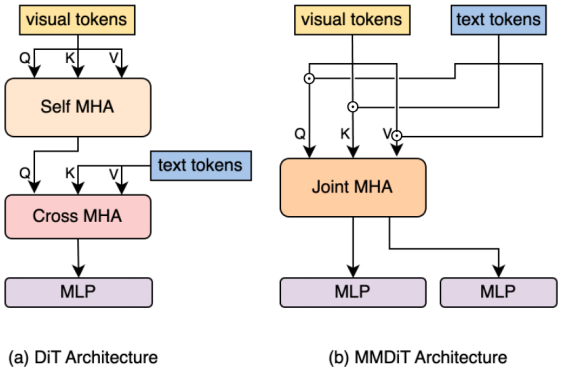

Figure 8. **DiT and MMDiT block architecture**. In MMDiT, after projections, visual and text tokens are concatenated for a joint self-attention.

Head Constraint Coefficient

We compared the generation results of attention sparsity under different constraint coefficients on the 1K image generation task of Stable Diffusion 3. We tested the SSIM and LPIPS of images generated with c=1 (fixed threshold for all heads), c=1.5, c=2, and without any constraints, against the original generated images. We found that the model achieved the lowest LPIPS (0.238) and highest SSIM (0.644) when the constraint coefficient was set to 1.5. Further relaxation of the constraint coefficient might lead to a decrease in similarity. Based on this finding, we set c=1.5 as the default constraint coefficient and used this coefficient to generate the experimental results presented in the other parts of this paper.

Table 5. SD3 1K Image Generation Similarity under Different Constraint Coefficient

Constraint Coefficient	Attention Sparsity	LPIPS	SSIM
-	0.50	0.249	0.640
$c = 1$	0.55	0.240	0.641
$c = 1.5$	0.55	**0.238**	**0.644**
$c = 2$	0.55	0.253	0.627

Computation Memory Analysis

Arrow attention can be seen as a block sparse attention, and the arithmetic intensity is $\frac{2N(1-r)}{2+(1-r)\times\frac{N}{B_r}}$. N, r, and B_r stand for sequence length, sparsity ratio, and CUDA block size. By roofline model analysis, for 2K image generation on A100, when $r \geq 0.95$, the attention computation is memory bound; otherwise, it is compute bound.

Extension to Video Generation

Our primary focus is high-resolution image generation. However, our framework is compatible with video generation models (e.g., CogViewX, HunyuanVideo, etc.). For video generation, beyond spatial redundancy, there exists additional redundancy, such as temporal redundancy. This redundancy can be effectively captured through techniques like token reordering that are discussed in recent video generation model acceleration works [46, 56, 60]. By adding the pattern arrow attention with token ordering, Our framework can be extended to support video generation compression better.

Perceptual Quality Examples

ImageReward [49] is a metric that reflects human preference on text-to-image tasks. Table 6 shows that our models has higher reward. Here we provide more generation results of our method in Fig 9, Fig 10, Fig 11

Table 6. SD3 1K ImageReward

Method	Reward	Rank
DiTFastAttn	0.242	5
DiTFastAttn w/ AA	0.255	4
Attention Caching	0.425	3
DiTFastAttnV2, $\delta = 0.4$	**0.487**	1
DiTFastAttnV2, $\delta = 0.6$	0.475	2

Theoretical Support

We posit that heterogeneity in transformers arises because heads operate independently and are initialized randomly, causing them to develop different functional roles. Recent work [59, 61] suggests that random initialization can be the reason behind attention heterogeneity. These work serves as theoretical support of attention head heterogeneity.

Figure 9. **SD 3 1K generation results**

Figure 10. **FLUX 1K generation results**

Figure 11. **FLUX 2K generation results**