# Efficient Diffusion Transformer with Step-wise Dynamic Attention Mediators

Yifan Pu[1*], Zhuofan Xia[1*], Jiayi Guo[1*], Dongchen Han[1], Qixiu Li[1], Duo Li[1], Yuhui Yuan[2], Ji Li[2], Yizeng Han[1], Shiji Song[1], Gao Huang[1(✉)], and Xiu Li[1(✉)]

[1] Tsinghua University, Beijing 100084, China
{puyf23, xzf23, guo-jy20}@mails.tsinghua.edu.cn
{shijis, gaohuang}@tsinghua.edu.cn
[2] Microsoft Research Asia
{yuhui.yuan, ji.li}@microsoft.com

**Abstract.** This paper identifies significant redundancy in the query-key interactions within self-attention mechanisms of diffusion transformer models, particularly during the early stages of denoising diffusion steps. In response to this observation, we present a novel diffusion transformer framework incorporating an additional set of mediator tokens to engage with queries and keys separately. By modulating the number of mediator tokens during the denoising generation phases, our model initiates the denoising process with a precise, non-ambiguous stage and gradually transitions to a phase enriched with detail. Concurrently, integrating mediator tokens simplifies the attention module's complexity to a linear scale, enhancing the efficiency of global attention processes. Additionally, we propose a time-step dynamic mediator token adjustment mechanism that further decreases the required computational FLOPs for generation, simultaneously facilitating the generation of high-quality images within the constraints of varied inference budgets. Extensive experiments demonstrate that the proposed method can improve the generated image quality while also reducing the inference cost of diffusion transformers. When integrated with the recent work SiT, our method achieves a state-of-the-art FID score of 2.01. The source code is available at https://github.com/LeapLabTHU/Attention-Mediators.

**Keywords:** Diffusion Transformer · Dynamic Neural Network

## 1 Introduction

Exhibiting unprecedented capabilities in the fields of language processing [1, 6, 14,62,70] and visual recognition [18,42,45,55,61], Transformers [71] have recently achieved remarkable performance in visual generation as backbones in diffusion models [5, 57]. The inherent simplicity, effectiveness, and scalability of these Diffusion Transformers (DiTs) position themselves as appealing alternatives to

∗ Equal contribution. ✉ Corresponding authors.

previously prominent U-Net structures [63–66], promoting the emergence of high-resolution and high-quality image/video generation applications, such as Stable Diffusion V3 [17], Pixart-$\alpha/\Sigma/\delta$ [9–11], Hunyuan-DiT [44] and Sora [5].

Despite the rapid progress of Diffusion Transformers, widespread criticism has arisen due to their substantial consumption of computing resources and the associated inference time overhead [11, 54, 83] resulting from the global attention mechanism. This obstacle impedes the practical deployment of Diffusion Transformers for large-scale client usage, particularly when dealing with high-resolution images [11, 50] and relatively long videos [48, 52]. While several works [12,19,89] have been proposed to accelerate the attention process in visual recognition tasks, this topic remains largely unexplored in the realm of visual generation. Therefore, it is crucial to develop an efficient Diffusion Transformer to address high resource consumption concerns and enhance overall usability.

In this paper, we expedite the diffusion generation process by leveraging the inherent structural redundancy [53, 69, 79, 88] in Diffusion Transformers across different denoising time steps. We start by identifying the redundancies in the query-key interaction process during the self-attention operation at each layer in Transformer diffusers. To analyze quantitatively, we design a Jensen–Shannon divergence-based metric to measure the query-key interaction redundancy, *i.e.*, comparing the attention distribution similarities among each query. We come up with two key findings: (1) Extensive query-key redundancy is evident in all of the self-attention layers, indicating many tokens would be homogeneous after self-attention; (2) The redundancy is particularly pronounced in the initial steps while gradually diminishing in the subsequent steps as denoising goes on, suggesting the fully one-to-one attention in the early steps be dispensable.

To fully take advantage of this redundancy, we introduce an extra set of tokens in the conventional self-attention layers, dubbed **attention mediators**, to streamline the interaction process between queries and keys, condensing the actual interactions in the attention between queries and keys. To be specific, the number of mediator tokens is set lower than that of queries and keys, *e.g.*, less than 10% of the original tokens. These mediator tokens first aggregate the information from keys with softmax attention, forming packed representations. Then, the compressed information is propagated to queries in another softmax attention as the final output. The abbreviated mediators bottleneck the attention and hence confine its redundancy, further reducing the computation cost via interchanging the attention computation order.

In addition to attention mediators, the redundancy variations across time steps elicit a new dynamic strategy for adjusting the number of mediator tokens at different time steps. Specifically, during the early steps where the redundancy is prominent, we utilize a smaller number of mediator tokens to reduce similar information aggregation effectively. When redundancy gradually diminishes during the later steps, we dynamically increase the number of mediator tokens to generate more detailed and diversified features. In practice, the schedule of switching mediators is determined by the samples' latent distance between each

pair of adjacent denoising steps. This dynamic strategy maintains mediator token efficiency while enhancing generation quality and diversity.

We evaluated our proposed method using the very recent SiT [51] model. Extensive experimental results demonstrate that our approach achieves superior generation quality (as indicated by a lower FID [32]) and reduces computational complexity (measured in FLOPs) during generation. When combined with the SiT-XL/2 model, our method achieves a state-of-the-art FID score.

## 2 Related Works

### 2.1 Diffusion Transformers

Recent advancements in diffusion models [2, 15, 21, 33, 46] have typically utilized the U-Net architecture [65]. However, a growing body of research [3, 57, 86] has begun to explore the potential of employing the Vision Transformer (ViT) [16] as an alternative backbone for such models. U-ViT [3] interprets various inputs (*e.g.*, time, conditions, and noisy image patches) as tokens while drawing inspiration from U-Net to implement skip connections between the model's shallow and deep layers. DiT [57] demonstrates the scalability of ViT for diffusion models, surpassing the performance of U-Net-based diffusion models on ImageNet. Building upon DiT, SiT [51] introduces an interpolant framework, moving from discrete to continuous time and exploring various diffusion coefficients, thereby achieving superior results. MaskDiT [89] pioneers the use of masked training to reduce the computational expense of training diffusion models. MDT [19] additionally proposes a masked latent modeling technique, and MDTv2 further refines this approach with a more efficient macro network architecture and training strategy, improving the FID and accelerating the learning process. HDiT [12] leverages transformers to devise a high-resolution training methodology that scales linearly with pixel count. FiT [89] conceptualizes images as sequences of dynamically sized tokens to generate images, facilitating image generation at varying resolutions and aspect ratios. These investigations confirm that transformer-based models are effective in visual generation tasks and can be scalable. Although these works have demonstrated the effectiveness of transformers in diffusion models and have further improved the FID or training speed by optimizing the diffusion structure or learning strategies, the inner design structure of the Diffusion Transformer backbone is still not well explored.

### 2.2 Attention with Linear Complexity

One line of works achieves linear computational complexity by restricting receptive fields, including Shifted-window attention [45], Neighborhood Attention [31]. These works bring locality back into the vision transformer architecture, while the global context awareness is somewhat affected. In contrast to the idea of restricting receptive fields, another line of researcgh directly uses linear attention to address the computational challenge by reducing computation complexity. The

pioneer work [39] discards the Softmax function and replaces it with a mapping function $\phi$ applied to $Q$ and $K$, thereby reducing the computation complexity to $\mathcal{O}(N)$. However, such approximations led to substantial performance degradation. To tackle this issue, Efficient Attention [68] applies the Softmax function to both $Q$ and $K$. SOFT [49] and Nyströmformer [82] employ matrix decomposition to further approximate Softmax operation. Castling-ViT [87] uses Softmax attention as an auxiliary training tool and fully employs linear attention during inference. FLatten Transformer [22] proposes a focused function and adopts depthwise convolution to promote feature diversity limited by linear operations.

Furthermore, Agent Attention [23] and Anchored Stripe Attention [43] introduce another group of tokens as the bridge between queries and keys, which is equivalent to linear attention, achieving favorable performance on recognition tasks and low-level visions, respectively. In this paper, we build our work upon this architecture and comprehend the extra group of tokens as semantically compressed information to guide the diffusion process to generate images.

### 2.3 Dynamic Neural Networks

In contrast to static models, which have fixed computational graphs and parameters at the inference stage, dynamic neural networks [25,76] can adapt their structures or parameters to different inputs, leading to notable advantages in terms of performance, adaptiveness [20, 85], computational efficiency [72, 84], and representational power [60]. Dynamic networks are typically categorized into three types: sample-wise [24,28,36,58,73,77,78], spatial-wise [26,27,29,37,56,74,80,81], and temporal-wise [30, 75]. Since the breakthrough query-based visual recognition model DETR [7], a new query-based dynamic network has begun to develop [59]. In this work, we introduce a novel temporal-wise dynamic approach. Contrary to the former works, which study the dynamic mechanism along the video time dimension, we explore the redundancy across the diffusion-denoising time steps in this paper. We dynamically change the number of mediator tokens, conditioned on the generation process of different image samples, and achieve better FID-50K results with less computational complexity.

## 3 Attention Redundancies Along Denoising Steps

In this section, we examine redundancies in conventional self-attention operations. Initially, we provide a brief overview of attention computation in Transformer architectures. Subsequently, we introduce a quantitative metric designed to analyze redundancies in query-key interactions. Our findings reveal that significant redundancies exist in Diffusion Transformers, and the extent of this redundancy decreases as the denoising procedure progresses.

### 3.1 Background of Attention

We first revisit the attention mechanism [71] in Diffusion Transformers [51, 57]. The latent Diffusion Transformer takes a latent token sequence $\boldsymbol{z}_{l-1} \in \mathbb{R}^{N \times C}$ from

the previous layer $l-1$ as input ($N$ is the token number and $C$ is the hidden dimension), then projects it into the query, key, and value sequences with three linear projection layers, denoted as $\mathbf{W_q}, \mathbf{W_k}, \mathbf{W_v} \in \mathbb{R}^{C\times C}$ (bias omitted):

$$\boldsymbol{q} = \boldsymbol{z}_{l-1}\mathbf{W_q},\ \ \boldsymbol{k} = \boldsymbol{z}_{l-1}\mathbf{W_k},\ \ \boldsymbol{v} = \boldsymbol{z}_{l-1}\mathbf{W_v}. \tag{1}$$

Then $\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v} \in \mathbb{R}^{N\times C}$ are divided into $M$ heads $\boldsymbol{q}^{(m)}, \boldsymbol{k}^{(m)}, \boldsymbol{v}^{(m)} \in \mathbb{R}^{N\times d}$ in terms of channel $C$, with head dimension of $d = C/M$. Within each head, the similarity of each query $\boldsymbol{q}^{(m)}$ and key $\boldsymbol{k}^{(m)}$ is computed as:

$$\mathbf{A}^{(m)} = \text{Softmax}\left(\boldsymbol{q}^{(m)}\boldsymbol{k}^{(m)\top}/\sqrt{d}\right), \tag{2}$$

where the attention map $\mathbf{A}^{(m)}$ is an $N \times N$ matrix containing elements in the range $[0, 1]$, and the sum of each row is normalized to 1. The attention mechanism reweights the value sequence according to the attention map, $\boldsymbol{h}^{(m)} = \mathbf{A}^{(m)}\boldsymbol{v}^{(m)}$, to dynamically adjust the outputs based on the dependency of each token in the inputs. In the end, each head of the reweighted representation is concatenated together to produce the final output of this layer $l$, written as:

$$\boldsymbol{z}_l = \text{Concat}\left(\boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}, \ldots, \boldsymbol{h}^{(M)}\right)\mathbf{W_O}, \tag{3}$$

where $\mathbf{W_O} \in \mathbb{R}^{C\times C}$ (bias omitted) is a linear projection layer to promote interaction between different heads in the multi-head attention layer.

We view each row of $\mathbf{A}^{(m)}$ in Eq. (2) as a probabilistic distribution between one query and all the keys, *e.g.*, the $i$-th row $\mathbf{A}_i^{(m)} \in \mathbb{R}^{1\times N}$ depicts how the $N$ key tokens contribute to the output of the $i$-th query token, on the $m$-th attention head. Since the output of $i$-th token $\boldsymbol{h}_i^{(m)} = \mathbf{A}_i^{(m)}\left(\boldsymbol{z}_{l-1}\mathbf{W_v}\right)^{(m)}$ only distinguishes other tokens by the distribution $\mathbf{A}_i^{(m)}$, the feature diversity in the output sequence of the attention is determined by this distribution. If different queries $\boldsymbol{q}_{i_1}$ and $\boldsymbol{q}_{i_2}$ $(i_1 \neq i_2)$ share similar probabilistic distributions over keys, *i.e.*, $\mathcal{D}\left(\mathbf{A}_{i_1}^{(m)}, \mathbf{A}_{i_2}^{(m)}\right) \approx 0$ for some distribution similarity metric $\mathcal{D}(\cdot,\cdot)$, the output $\boldsymbol{h}_{i_1}^{(m)}$ and $\boldsymbol{h}_{i_2}^{(m)}$ would be rather close, leading to redundant representations and a lack of spatial diversity in the diffusion noise prediction process.

### 3.2 Jensen-Shannon Divergence as A Redundancy Metric

We adopt Jensen-Shannon Divergence (JSD) as the redundancy metric $\mathcal{D}$ to study the spatial redundancy in attention on latent tokens quantitatively. JSD is a symmetric divergence that combines two Kullback–Leibler Divergence (KLD). Given two probabilistic distributions $\mathbb{P}_1(X)$ and $\mathbb{P}_2(X)$ in which $X$ is a discrete random variable with $K$ possible values, the KLD is defined as

$$\mathcal{D}_{\text{KL}}\left(\mathbb{P}_1 \| \mathbb{P}_2\right) = \sum_{k=1}^{K} \mathbb{P}_1(X=k)\left[\ln \mathbb{P}_1(X=k) - \ln \mathbb{P}_2(X=k)\right]. \tag{4}$$

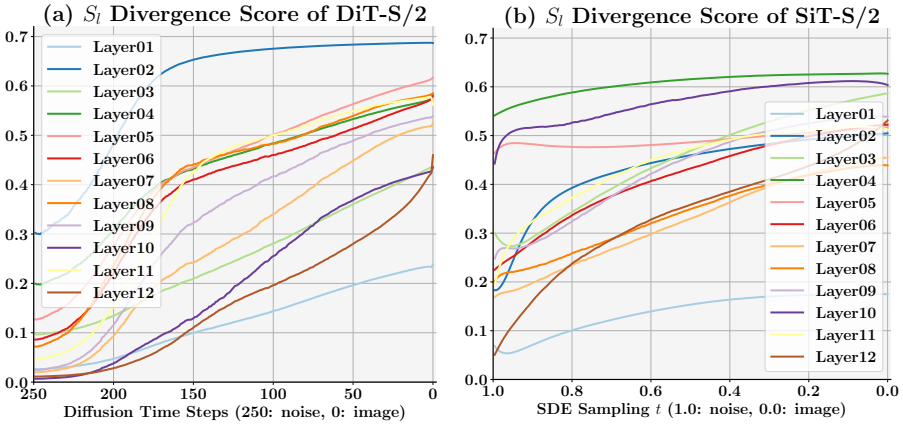


**Fig. 1:** (a) shows the JSD-based redundancy score defined in Sec. 3.2 evaluated on DiT-S/2 model along with diffusion time steps. The score is computed over 32 samples and averaged by different attention heads in every layer. (b) shows the same redundancy score of all the 12 layers of SiT-S/2 model with the SDE sampler.

Then the JSD is defined with a mixture distribution $\mathbb{M} = \frac{1}{2}(\mathbb{P}_1 + \mathbb{P}_2)$, by averaging the KLD of $\mathbb{P}_1$ from $\mathbb{M}$ and the KLD of $\mathbb{P}_2$ from $\mathbb{M}$, written as

$$\mathcal{D}_{\mathrm{JS}}(\mathbb{P}_1 \| \mathbb{P}_2) = \frac{1}{2}\left[\mathcal{D}_{\mathrm{KL}}(\mathbb{P}_1 \| \mathbb{M}) + \mathcal{D}_{\mathrm{KL}}(\mathbb{P}_2 \| \mathbb{M})\right]. \tag{5}$$

The JSD is symmetric and bounded in that $\mathcal{D}_{\mathrm{JS}}(\mathbb{P}_1 \| \mathbb{P}_2) = 0$ when $\mathbb{P}_1$ and $\mathbb{P}_2$ are identical, and $\mathcal{D}_{\mathrm{JS}}(\mathbb{P}_1 \| \mathbb{P}_2) \to \ln 2$ when the support of $\mathbb{P}_1$ and $\mathbb{P}_2$ are disjoint. JSD decreases as two distributions are closer and increases vice versa.

For the query token sequence, we compare the attention distribution by each pair of queries using Jensen-Shannon Divergence and then accumulate the divergence to each query token as the final redundancy score metric, which we define as follows for the $l$-th layer in the Diffusion Transformers:

$$S_l = \frac{2}{MN(N-1)} \sum_{m=1}^{M} \sum_{i_1=1}^{N-1} \sum_{i_2=i_1+1}^{N} \mathcal{D}_{\mathrm{JS}}\left(\mathbf{A}_{i_1}^{(m)}, \mathbf{A}_{i_2}^{(m)}\right). \tag{6}$$

This score computes the JSD of every attention distribution pair in the latent token sequence and reduces over $\frac{N(N-1)}{2}$ pairs and $M$ attention heads. A *high* $S_l$ means that the averaged attention maps among the tokens are in *low* similarity in the $l$-th layer, indicating a *low* spatial redundancy. On the contrary, a low $S_l$ means the redundancy in the $l$-th layer is relatively high.

### 3.3 Redundancies Along Time Steps

We measure the $S_l$ of both the DiT-S/2 [57] and SiT-S/2 model [51]. We randomly sample 512 images with a pretrained model and record the $S_l$ of all the

diffusion transformer layers and all the denoising time steps (all the SDE sampling $t$ in SiT). The results for DiT-S/2 and SiT-S/2 are illustrated in Fig. 1(a) and Fig. 1(b), respectively. Notably, the redundancy of self-attention is *inversely* proportional to $S_l$. As a result, we get two observations from Fig. 1. First, massive query-key redundancy exists in the attention operation of the diffusion transformers. For example, in some layers (*e.g.* layer 10 in DiT-S/2), the inner-query distance is nearly zero in the first several time steps, implying that almost all the queries are akin and redundant. The second observation is that redundancy gradually decreases as the denoising process continues. It is implied that the queries become more diverse in the latter denoising steps.

Based on the above phenomenon, we design mediator tokens that interact with query and key tokens separately, thus compressing the excessive attention between queries and keys. The number of mediator tokens can be adjusted in different time steps, thus adapting the different degrees of redundancy inside different phases of the denoising process. We present the detailed explanation of our method in the following section.

## 4 Efficient DiTs with Attention Mediators

In this section, we introduce the attention mediator mechanism to leverage the redundancy efficiently in Sec. 4.1, building up a dynamic architecture of Diffusion Transformer. To further boost the efficiency of Dynamic Diffusion Transformers, we devise an algorithm in Sec. 4.3 to speed up the sampling process and fit the computational budgets via dynamically adjusting mediator tokens.

### 4.1 Attention Mediators

We present the attention mediators to regulate the attention between every two query and key pairs. The high-level idea of attention mediators is to use an additional group of tokens to compress the interaction between the queries and keys. The additional tokens, which we name it as mediator tokens, usually have a smaller number than queries or keys, serving as a condensed supervisor over the attention interaction. We present the detail as follows.

In each head of the multi-head attention module, besides the query $\boldsymbol{q}^{(m)}$, key $\boldsymbol{k}^{(m)}$, and value $\boldsymbol{v}^{(m)}$ tokens, we introduce a set of mediator tokens $\boldsymbol{t}^{(m)} \in \mathbb{R}^{n \times d}$, where $n$ is the mediator token length and $n \ll N$. The mediator tokens first interact with the key tokens to get the intermediate result $\boldsymbol{v}_{\text{med}}^{(m)}$:

$$\boldsymbol{v}_{\text{med}}^{(m)} = \text{Softmax}\left(\boldsymbol{t}^{(m)}\boldsymbol{k}^{(m)\top}/\sqrt{d}\right)\boldsymbol{v}^{(m)}, \tag{7}$$

where $\boldsymbol{v}_{\text{med}}^{(m)} \in \mathbb{R}^{n \times d}$. Then the mediator token interacts with the query tokens and extracts the results from the intermediate result $\boldsymbol{v}_{\text{med}}^{(m)}$:

$$\boldsymbol{h}^{(m)} = \text{Softmax}\left(\boldsymbol{q}^{(m)}\boldsymbol{t}^{(m)\top}/\sqrt{d}\right)\boldsymbol{v}_{\text{med}}^{(m)}. \tag{8}$$

In this way, a condensed set of mediator tokens interacts with the queries and keys separately, avoiding redundancy when they interact indirectly.

The mediator tokens are obtained by adaptively pooling the query tokens into a small number of tokens. Considering the noise predicted by the transformer has spatial structured information, we first reshape the query tokens into the latent image shape $\mathbb{R}^{H\times W\times d}$ and then pool it in the spatial dimensions to get $\mathbb{R}^{h\times w\times d}$. The pooled queries are finally reshaped to the mediator tokens $\boldsymbol{t}^{(n)} \in \mathbb{R}^{n\times d}$, where $n \ll N$ because $(h \times w) \ll (H \times W)$.

### 4.2 Complexity Analysis

It is noteworthy that by incorporating an additional, compact set of tokens, we achieve a reduction in redundancy within the attention mechanism. Simultaneously, the computational complexity inherent to the attention operation is diminished. We provide the subsequent analysis.

We begin by mixing and combining Eq. (7) and Eq. (8) to formulate the final output of self-attention with mediator tokens:

$$\boldsymbol{h}^{(m)} = \underbrace{\text{Softmax}\left(\boldsymbol{q}^{(m)}\boldsymbol{t}^{(m)\top}/\sqrt{d}\right)\underbrace{\text{Softmax}\left(\boldsymbol{t}^{(m)}\boldsymbol{k}^{(m)\top}/\sqrt{d}\right)\boldsymbol{v}^{(m)}}_{\text{Step 1: } \mathbb{R}^{n\times N}\cdot\, \mathbb{R}^{N\times d}\,\rightarrow\, \mathcal{O}(Nnd)}}_{\text{Step 2: } \mathbb{R}^{N\times n}\cdot\, \mathbb{R}^{n\times d}\,\rightarrow\, \mathcal{O}(Nnd)}. \tag{9}$$

Since queries $\boldsymbol{q}^{(m)}$ and keys $\boldsymbol{k}^{(m)}$ are decoupled by the mediators, we can interchange the computation order of the queries, keys and values in attention. Unlike previous vanilla self-attention that firstly computes $\boldsymbol{q}^{(m)}$ and $\boldsymbol{k}^{(m)}$, we first aggregate values $\boldsymbol{v}^{(m)}$ with precomputed $\boldsymbol{A}_{\text{tk}}^{(m)} = \text{Softmax}\left(\boldsymbol{t}^{(m)}\boldsymbol{k}^{(m)\top}/\sqrt{d}\right)$, as shown in Step 1 of Eq. (9). The complexity of step 1 in multiplying an $n \times N$ matrix and an $N \times d$ matrix is $\mathcal{O}(Nnd)$, as well as computing $\boldsymbol{A}_{\text{tk}}^{(m)}$, which involves multiplying an $n \times d$ matrix and an $N \times d$ matrix. Thus, the overall complexity of Step 1 is no more than $2Nnd$, also controlled by $\mathcal{O}(Nnd)$. The result of Step 1 has the shape of $\mathbb{R}^{n\times d}$, therefore the information propagation to queries of step 2 with $\boldsymbol{A}_{\text{qt}}^{(m)} = \text{Softmax}\left(\boldsymbol{q}^{(m)}\boldsymbol{t}^{(m)\top}/\sqrt{d}\right)$ is also an $\mathcal{O}(Nnd)$ complex operation.

To summarize, both Steps 1 and 2 in Eq. (9) have $\mathcal{O}(Nnd)$ complexity, with $N$ latent tokens, $n$ mediator tokens, and $d$ feature dimensions in each attention head. The proposed attention module achieves linear complexity relative to $N$, $n$, and $d$. Summing all heads together, the proposed mediator attention has an $\mathcal{O}(nNC)$ complexity. Compared with the vanilla self-attention, which directly multiplies queries and keys together to aggregate values and get $\mathcal{O}(N^2C)$ complexity, our method significantly reduces computational demands, given that the mediator token count $n$, is significantly less than the image token count $N$. To compensate the potential loss of feature diversity in linear complexity attention, we adopt a depthwise convolution following Flatten transformer [22].

### 4.3 Time Step-wise Mediator Adjusting

Fig. 1 illustrates the variation in attention redundancy across different diffusion denoising time steps, revealing a gradual decrease in redundancy throughout the process. Understanding the attention mediator tokens as a means of compressing tokens between query and value tokens, we exploit this phenomenon, as shown in Fig. 1, to dynamically adjust the number of mediator tokens, increasing them from loss to more along the diffusion denoising steps.

Given the variability of the denoising procedure across image samples, we introduce a sample-specific method for dynamically adjusting the number of mediator tokens. This approach allows for a customized mediator token adjustment schedule for each sample, based on its unique denoising process.

To quantify the changes in latent features between adjacent time steps, we calculate the distance between each pair of subsequent time steps, denoted as $\Delta_t = \|x_t - x_{t+1}\|$, alongside recording the initial denoising difference $\Delta_0 = \|x_0 - x_1\|$. The denoising process begins with a Diffusion Transformer featuring a smaller number $n_1$ of mediator tokens. Upon the latent difference falling below a threshold $\rho_0$ of the initial difference $\Delta_0$, we transition to a Diffusion Transformer with an increased number $n_2$ of mediator tokens.

$$n_t = \begin{cases} n_1, \Delta_t > \rho_0 \cdot \Delta_0, \\ n_2, \Delta_t \leq \rho_0 \cdot \Delta_0. \end{cases} \tag{10}$$

This process is further refined by introducing additional thresholds for change, utilizing varying numbers of mediator tokens at each stage:

$$n_t = \begin{cases} n_1, \Delta_{\mathrm{t}} > \rho_0 \cdot \Delta_0, \\ n_2, \Delta_{\mathrm{t}} \leq \rho_1 \cdot \Delta_0, \\ \quad \vdots \\ n_{\mathrm{k}}, \Delta_{\mathrm{t}} \leq \rho_{\mathrm{k}-1} \cdot \Delta_0. \end{cases} \tag{11}$$

## 5 Experiments

In this section, we empirically evaluate the proposed sample-wise adaptive mediator tokens adjustment method on the state-of-the-art diffusion transformer SiT [51]. We begin by introducing the experiment settings in Sec. 5.1, which include the dataset description and training hyper-parameters. The experiment results for different numbers of mediator tokens are presented in Sec. 5.2. In Sec. 5.3, we show how to optimize the schedule for adjusting the mediator tokens. Then, the effectiveness of the time step-wise mediator adjustment mechanism on larger models and higher resolutions is demonstrated in Sec. 5.4. We also compare our method with some state-of-the-art approaches in Sec. 5.5. Finally, more ablation studies regarding our method and the generation visualization results are presented in Sec. 5.6 and Sec. 5.7, respectively.

**Table 1:** Effectiveness of static mediator tokens. $n$ is the mediator tokens number.

| Model | FLOPs(G) | FID (↓) | sFID (↓) | IS (↑) | Precision (↑) | Recall (↑) |
|---|---|---|---|---|---|---|
| SiT-S/2 (baseline) | 6.06 | 58.61 | 9.25 | 24.31 | 0.41 | 0.59 |
| + Ours ($n = 4$) | **5.49** | 57.67 | 10.01 | 26.66 | 0.42 | 0.56 |
| + Ours ($n = 16$) | 5.55 | 54.55 | 9.28 | 26.55 | **0.43** | 0.59 |
| + Ours ($n = 64$) | 5.78 | **53.57** | **9.01** | **27.26** | **0.43** | **0.61** |

### 5.1 Experimental Setups

Following DiT [57] and SiT [51], we train class-conditional diffusion transformer models on the highly-competitive generative modeling benchmark ImageNet-1k [13]. We adopt AdamW [40, 47] optimizer to train all the diffusion models with no weight decay. For $256 \times 256$ image resolution models, we train them from scratch with a global batch size of 256 for 400K iterations. The global learning rate is set as constant $1 \times 10^{-4}$ during all training steps. We only use simple random horizontal flops data augmentation and maintain an exponential moving average (EMA) of the model weights over training with a decay of 0.9999.

### 5.2 Effectiveness of Attention Mediator Tokens

To verify the effectiveness of the proposed mediators, we replace the standard self-attention layers in SiT-S/2 [51] with the mediator-token ones. The experiments are conducted at a $256 \times 256$ resolution, and the images are sampled without using classifier-free guidance. Tab. 1 shows the results for different numbers of static mediator tokens, which means the token number is static across different denoising time steps. It is observed that by compressing the query-key interaction process, our method not only reduces the computational complexity in FLOPs but also achieves a higher generated image quality in FID.

### 5.3 Exploring Optimized Mediator Token Adjustment Schedule

Since determining optimized thresholds ($\rho_i$ in Eq. (11)) is non-trivial, we conduct a small-scale grid search to explore reasonable mediator token number change thresholds. Specifically, we use the three models introduced in Tab. 1. We sweep the first threshold $\rho_0$ in $\{1.0, 0.9, \cdots, 0.1, 0.0\}$, and sweep the second threshold $\rho_1$ in $\{\rho_0, \rho_0 - 0.1, \cdots, 0.1, 0.0\}$. In this way, this search space not only includes the ensemble of these three models with different numbers of mediator tokens, but also contains two-model ensembles and a single model. The choice of distance function, as described in Sec. 4.3, is also ablated between L1 and L2 distance.

The results regarding the trade-off between FID/sFID-50K and computation cost in GFLOPs are illustrated in Fig. 2(a) and Fig. 2(b). We plot all the results under different thresholds, along with their envelope curves. The thresholds in the envelope curves are considered optimized. We also compare the effectiveness of using L1 versus L2 distance and find that the L1 distance is the better choice.

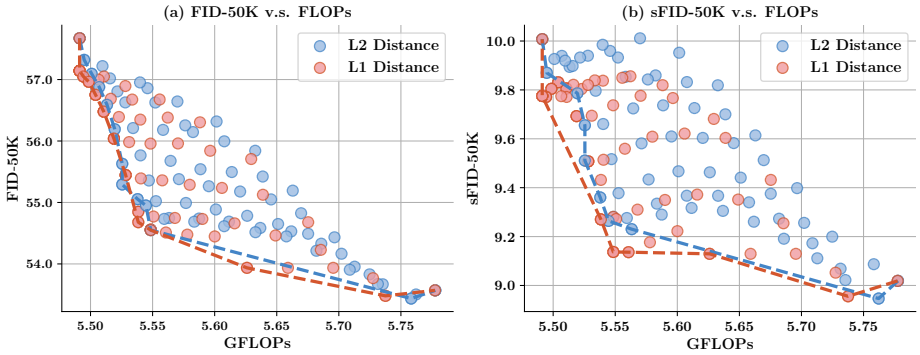


**Fig. 2:** Ablation for optimized mediator token adjustment schedule. (a) Trade-off between FID-50K and FLOPs. (b) Trade-off between sFID-50K and FLOPs.

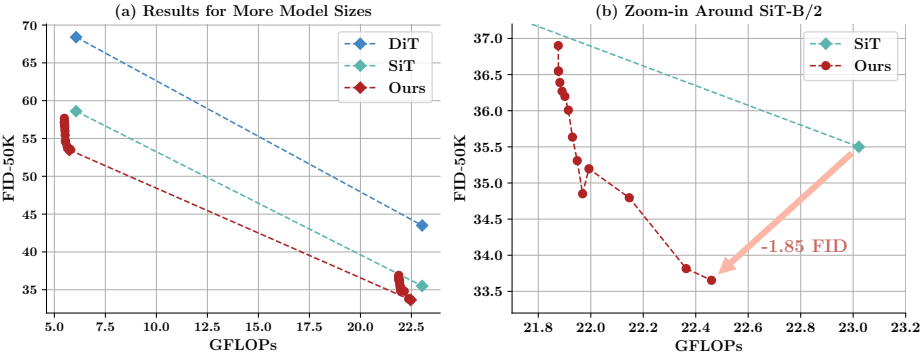


**Fig. 3:** Main Results of the proposed method in $256 \times 256$ resolution. Each string of red dots is obtained by adjusting the mediator token number with optimized thresholds. (a) Comparison with DiT [57] and SiT [51]; (b) Zoomed in results around SiT-B/2.

### 5.4 Main Results

We adopt the optimized thresholds obtained in Sec. 5.3 and repeat the aforementioned experiment on a larger scale model SiT-B/2. The results in Fig. 3 show that our method consistently outperform both DiT and SiT (Fig. 3 (a)) and this phenomenon is consistent between different model sizes (Fig. 2 (a) for SiT-S/2, Fig. 3 (b) for SiT-B/2). Specifically, our method can get a better FID score (1.85 lower than SiT-B/2) with even less computation budget.

We further conduct experiment on generating higher resolution images. The $512 \times 512$ resolution models are finetuned from $256 \times 256$ models with a global batch size of 64 for 400K iterations, while $1024 \times 1024$ models are finetuned from $512 \times 512$ counterparts with a global batch size of 16 for 400K iterations. For testing $512 \times 512$ resolution models, we generate 10K images with our model and compute the FID with 512 resolution reference batch obtained from

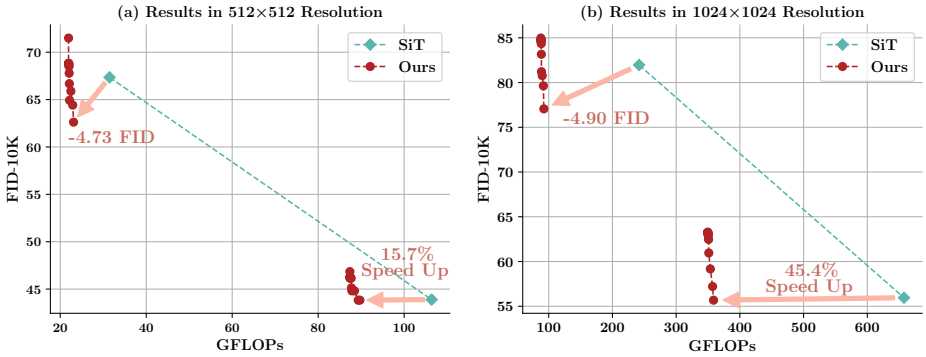


**Fig. 4:** High resolution image generation results.

**Table 2:** Benchmarking class-conditional image generation on ImageNet 256×256.

| Model | FID↓ | sFID↓ | IS↑ | Precision↑ | Recall↑ |
|---|---|---|---|---|---|
| BigGAN-deep [4] | 6.95 | 7.36 | 171.4 | **0.87** | 0.28 |
| StyleGAN-XL [67] | 2.30 | **4.02** | 265.12 | 0.78 | 0.53 |
| Mask-GIT [8] | 6.18 | - | 182.1 | - | - |
| ADM [15] | 10.94 | 6.02 | 100.98 | 0.69 | **0.63** |
| ADM-G, ADM-U | 3.94 | 6.14 | 215.84 | 0.83 | 0.53 |
| CDM [34] | 4.88 | - | 158.71 | - | - |
| RIN [38] | 3.42 | - | 182.0 | - | - |
| Simple Diffusion(U-Net) [35] | 3.76 | - | 171.6 | - | - |
| Simple Diffusion(U-ViT, L) | 2.77 | - | 211.8 | - | - |
| VDM++ [41] | 2.12 | - | 267.7 | - | - |
| DiT-XL$_{(\text{cfg}=1.5)}$ [57] | 2.27 | 4.60 | **278.24** | 0.83 | 0.57 |
| SiT-XL$_{(\text{cfg}=1.5)}$ [51] | 2.06 | 4.50 | 270.27 | 0.82 | 0.59 |
| Ours$_{(\text{cfg}=1.5)}$ | **2.01** | 4.49 | 271.04 | 0.82 | 0.60 |

`guided-diffusion`[3]. For 1024 × 1024 models, we randomly select 10K images from ImageNet validation set, resize them into $1024^2$ resolution, and compute FID ( with `clean-fid`[4] toolkit) with 10K images sampled by our model.

The high-resolution results is illustrated in Fig. 4, where we can find that: (1) the proposed method can still achieve better generated image quality (*e.g.*, for SiT-S/2, −4.90 FID for $1024^2$) with far fewer FLOPs, and (2) the speedup is even more significant as the image resolution increases (*e.g.*, for SiT-B/2, the speed-up increase from 15.7% in $512^2$ resolution to 45.4% in $1024^2$ resolution). This is because as the image resolution grows, the sequence length the attention operation needs to process also increases. At this point, the superiority of

[3] https://github.com/openai/guided-diffusion/tree/main/evaluations
[4] https://github.com/GaParmar/clean-fid

the linear complexity in our method becomes far more prominent compared to standard attention, which has quadratic complexity w.r.t the sequence length.

### 5.5 Comparsion with State-of-the-art

We compare our method against state-of-the-art class-conditional generative models with the highest complexity SiT-XL/2 model endowed with our method. We replace the first four self-attention layers with the proposed attention with mediator tokens, and finetune the modified model for 400K iterations. The results reported in Tab. 2 illustrate that when using classifier-free guidance (cfg=1.5), following the practice in DiT and SiT, our method outperforms all the prior diffusion models, achieving a remarkable FID-50K of 2.01.

### 5.6 Ablation Studies

**Table 3:** Effectiveness of static mediator tokens. $n$ is the mediator tokens number.

| Model | FLOPs(G) | FID (↓) | sFID (↓) | IS (↑) | Precision (↑) | Recall (↑) |
|---|---|---|---|---|---|---|
| SiT-S/2 (baseline) | 6.06 | 58.61 | 9.25 | 24.31 | 0.41 | 0.59 |
| $r = 0.875$ | 5.91 | 58.98 | 9.10 | 24.13 | 0.40 | 0.60 |
| $r = 0.750$ | 5.76 | 59.18 | 9.26 | 24.03 | 0.39 | 0.59 |
| $r = 0.625$ | 5.61 | 60.30 | 9.58 | 23.74 | 0.39 | 0.59 |
| $r = 0.500$ | 5.46 | 60.02 | 9.43 | 24.01 | 0.40 | 0.57 |
| Ours ($n = 64$) | 5.78 | 53.57 | 9.01 | 27.26 | 0.43 | 0.61 |

**Comparison with vanilla Q-K compression.** In order to verify that the proposed mediator token method is an effective way to leverage the query-key interaction redundancy, we design experiments where queries and keys are reduced in a simpler way. Specifically, in each self-attention layer of the SiT model, we modify the $\mathbf{W_q}$ and $\mathbf{W_k}$ linear projections from $\mathbb{R}^{C\times C}$ to $\mathbb{R}^{C\times rC}$ (where $r < 1$) dimensions. In this way, queries and keys also interact in a compressed space. We train this model with the same training recipe as SiT. The results in Tab. 3 show that although directly reducing the hidden dimension of queries and keys can save computation cost, the generated image quality drops dramatically. In contrast, the proposed method can increase the generated image quality as well as reduce the inference cost, verifying that our method is an effective way to leverage the redundancy in diffusion transformers.

**Fig. 5:** Sampled images by SiT-XL/2 models endowed with our method trained on ImageNet 256×256 resolution with cfg=4.0.

### 5.7 Visualization Results

In order to verify the proposed time step-wise dynamic mediator token adjusting token mechanism does not achieve a better numerical result by over-fitting the FID-50K metric, we visualize the sample images using the largest SiT-XL/2 based model. Following the common practice in the DiT [57] and the SiT [51], we set the classifier-free guidance as 4.0 to sample the images. The sampled results are visualized in Fig. 5, from which we can find that our method not only can achieve lower FID metric but also can generate high-quality images.

## 6 Conclusion

This paper proposed a novel diffusion transformer architecture in which an extra group of mediator tokens interact with the query tokens and key tokens separately, compressing the redundant query-key interaction during the denoising generation process. The number of mediator tokens adjusts across different denoising time steps conditioned on the difference between every two adjacent latent features in a simple-wise dynamic manner. Extensive quantitative experiments and qualitative generated results demonstrate the effectiveness of our method in alleviating attention redundancy and improving the generated image quality. Our method also reduces the computation complexity in the attention model since the proposed mechanism makes the attention operation have linear complexity with regard to the image token length.

## Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Grants 62321005 and 42327901.

## References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Aditya, R., Prafulla, D., Alex, N., Casey, C., Mark, C.: Hierarchical text-conditional image generation with clip latents. arXiv:2204.06125 (2022)
3. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: All are worth words: A vit backbone for diffusion models. In: IEEE CVPR (2023)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. In: ICLR (2019)
5. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), `https://openai.com/research/video-generation-models-as-world-simulators`
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS (2020)
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020)
8. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: IEEE CVPR (2022)
9. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-$\sigma$: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In: ECCV (2024)
10. Chen, J., Wu, Y., Luo, S., Xie, E., Paul, S., Luo, P., Zhao, H., Li, Z.: Pixart-$\delta$: Fast and controllable image generation with latent consistency models. In: ICML (2024)
11. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-$\alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: ICLR (2024)
12. Crowson, K., Baumann, S.A., Birch, A., Abraham, T.M., Kaplan, D.Z., Shippole, E.: Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In: ICML (2024)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE CVPR (2009)
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: ACL (2019)
15. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)

17. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Muller, J., Saini, H., and¨Dominik Lorenz, Y.L., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis (2024), https://stabilityai-public-packages.s3.us-west-2.amazonaws.com/Stable+Diffusion+3+Paper.pdf
18. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: EVA: Exploring the limits of masked visual representation learning at scale. In: IEEE CVPR (2023)
19. Gao, S., Zhou, P., Cheng, M.M., Yan, S.: Masked diffusion transformer is a strong image synthesizer. In: IEEE ICCV (2023)
20. Guo, J., Wang, C., Wu, Y., Zhang, E., Wang, K., Xu, X., Shi, H., Huang, G., Song, S.: Zero-shot generative model adaptation via image-specific prompt learning. In: IEEE CVPR (2023)
21. Guo, J., Xu, X., Pu, Y., Ni, Z., Wang, C., Vasu, M., Song, S., Huang, G., Shi, H.: Smooth diffusion: Crafting smooth latent spaces in diffusion models. In: IEEE CVPR (2024)
22. Han, D., Pan, X., Han, Y., Song, S., Huang, G.: FLatten transformer: Vision transformer using focused linear attention. In: IEEE ICCV (2023)
23. Han, D., Ye, T., Han, Y., Xia, Z., Song, S., Huang, G.: Agent attention: On the integration of softmax and linear attention. In: ECCV (2024)
24. Han, Y., Han, D., Liu, Z., Wang, Y., Pan, X., Pu, Y., Deng, C., Feng, J., Song, S., Huang, G.: Dynamic perceiver for efficient visual recognition. In: IEEE ICCV (2023)
25. Han, Y., Huang, G., Song, S., Yang, L., Wang, H., Wang, Y.: Dynamic neural networks: A survey. IEEE TPAMI (2021)
26. Han, Y., Huang, G., Song, S., Yang, L., Zhang, Y., Jiang, H.: Spatially adaptive feature refinement for efficient inference. IEEE TIP (2021)
27. Han, Y., Liu, Z., Yuan, Z., Pu, Y., Wang, C., Song, S., Huang, G.: Latency-aware unified dynamic networks for efficient image recognition. IEEE TPAMI (2024)
28. Han, Y., Pu, Y., Lai, Z., Wang, C., Song, S., Cao, J., Huang, W., Deng, C., Huang, G.: Learning to weight samples for dynamic early-exiting networks. In: ECCV (2022)
29. Han, Y., Yuan, Z., Pu, Y., Xue, C., Song, S., Sun, G., Huang, G.: Latency-aware spatial-wise dynamic networks. In: NeurIPS (2022)
30. Hansen, C., Hansen, C., Alstrup, S., Simonsen, J.G., Lioma, C.: Neural speed reading with structural-jump-lstm. In: ICLR (2019)
31. Hassani, A., Walton, S., Li, J., Li, S., Shi, H.: Neighborhood attention transformer. In: IEEE CVPR (2023)
32. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
33. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
34. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. JMLR (2022)
35. Hoogeboom, E., Heek, J., Salimans, T.: simple diffusion: End-to-end diffusion for high resolution images. In: ICML (2023)
36. Huang, G., Chen, D., Li, T., Wu, F., Van Der Maaten, L., Weinberger, K.Q.: Multi-scale dense networks for resource efficient image classification. In: ICLR (2018)
37. Huang, G., Wang, Y., Lv, K., Jiang, H., Huang, W., Qi, P., Song, S.: Glance and focus networks for dynamic visual recognition. IEEE TPAMI (2022)

38. Jabri, A., Fleet, D., Chen, T.: Scalable adaptive computation for iterative generation. In: ICML (2023)
39. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: ICML (2020)
40. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
41. Kingma, D.P., Gao, R.: Understanding the diffusion objective as a weighted integral of elbos. In: NeurIPS (2023)
42. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: IEEE ICCV (2023)
43. Li, Y., Fan, Y., Xiang, X., Demandolx, D., Ranjan, R., Timofte, R., Van Gool, L.: Efficient and explicit modelling of image hierarchies for image restoration. In: IEEE CVPR (2023)
44. Li, Z., Zhang, J., Lin, Q., Xiong, J., Long, Y., Deng, X., Zhang, Y., Liu, X., Huang, M., Xiao, Z., et al.: Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748 (2024)
45. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE ICCV (2021)
46. Liu, Z., Schaldenbrand, P., Okogwu, B.C., Peng, W., Yun, Y., Hundt, A., Kim, J., Oh, J.: Scoft: Self-contrastive fine-tuning for equitable image generation. In: CVPR (2024)
47. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
48. Lu, H., Yang, G., Fei, N., Huo, Y., Lu, Z., Luo, P., Ding, M.: VDT: General-purpose video diffusion transformers via mask modeling. In: ICLR (2023)
49. Lu, J., Yao, J., Zhang, J., Zhu, X., Xu, H., Gao, W., Xu, C., Xiang, T., Zhang, L.: Soft: Softmax-free transformer with linear complexity. In: NeurIPS (2021)
50. Lu, Z., Wang, Z., Huang, D., Wu, C., Liu, X., Ouyang, W., Bai, L.: FiT: Flexible vision transformer for diffusion model. In: ICML (2024)
51. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: ECCV (2024)
52. Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024)
53. Michel, P., Levy, O., Neubig, G.: Are sixteen heads really better than one? In: NeurIPS (2019)
54. Mo, S., Xie, E., Chu, R., Hong, L., Niessner, M., Li, Z.: DiT-3D: Exploring plain diffusion transformers for 3d shape generation. In: NeurIPS (2023)
55. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. TMLR (2024)
56. Pan, X., Ye, T., Xia, Z., Song, S., Huang, G.: Slide-transformer: Hierarchical vision transformer with local self-attention. In: IEEE CVPR (2023)
57. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: IEEE ICCV (2023)
58. Pu, Y., Han, Y., Wang, Y., Feng, J., Deng, C., Huang, G.: Fine-grained recognition with learnable semantic data augmentation. IEEE TIP (2023)

59. Pu, Y., Liang, W., Hao, Y., Yuan, Y., Yang, Y., Zhang, C., Hu, H., Huang, G.: Rank-detr for high quality object detection. In: NeurIPS (2024)
60. Pu, Y., Wang, Y., Xia, Z., Han, Y., Wang, Y., Gan, W., Wang, Z., Song, S., Huang, G.: Adaptive rotated convolution for rotated object detection. In: IEEE ICCV (2023)
61. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
62. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR (2020)
63. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML (2021)
64. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE CVPR (2022)
65. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
66. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)
67. Sauer, A., Schwarz, K., Geiger, A.: Stylegan-xl: Scaling stylegan to large diverse datasets. In: SIGGRAPH (2022)
68. Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient attention: Attention with linear complexities. In: WACV (2021)
69. Song, L., Zhang, S., Liu, S., Li, Z., He, X., Sun, H., Sun, J., Zheng, N.: Dynamic grained encoder for vision transformers. In: NeurIPS (2021)
70. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
71. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
72. Wang, C., Yang, Q., Huang, R., Song, S., Huang, G.: Efficient knowledge distillation from model checkpoints. In: NeurIPS (2022)
73. Wang, J., Pu, Y., Han, Y., Guo, J., Wang, Y., Li, X., Huang, G.: Gra: Detecting oriented objects through group-wise rotating and attention. In: ECCV (2024)
74. Wang, S., Wu, L., Cui, L., Shen, Y.: Glancing at the patch: Anomaly localization with global and local feature comparison. In: IEEE CVPR (2021)
75. Wang, Y., Chen, Z., Jiang, H., Song, S., Han, Y., Huang, G.: Adaptive focus for efficient video recognition. In: IEEE ICCV (2021)
76. Wang, Y., Han, Y., Wang, C., Song, S., Tian, Q., Huang, G.: Computation-efficient deep learning for computer vision: A survey. Cybernetics and Intelligence (2023)
77. Wang, Y., Huang, R., Song, S., Huang, Z., Huang, G.: Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. In: NeurIPS (2021)
78. Xia, Z., Han, D., Han, Y., Pan, X., Song, S., Huang, G.: Gsva: Generalized segmentation via multimodal large language models. In: IEEE CVPR (2024)
79. Xia, Z., Pan, X., Jin, X., He, Y., Xue', H., Song, S., Huang, G.: Budgeted training for vision transformer. In: ICLR (2023)
80. Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Vision transformer with deformable attention. In: IEEE CVPR (2022)

81. Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Dat++: Spatially dynamic vision transformer with deformable attention. arXiv preprint arXiv:2309.01430 (2023)
82. Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., Singh, V.: Nyströmformer: A nyström-based algorithm for approximating self-attention. In: AAAI (2021)
83. Xue, S., Yi, M., Luo, W., Zhang, S., Sun, J., Li, Z., Ma, Z.M.: SA-Solver: Stochastic adams solver for fast sampling of diffusion models. In: NeurIPS (2023)
84. Yang, Q., Wang, S., Lin, M.G., Song, S., Huang, G.: Boosting offline reinforcement learning with action preference query. In: ICML (2023)
85. Yang, Q., Wang, S., Zhang, Q., Huang, G., Song, S.: Hundreds guide millions: Adaptive offline reinforcement learning with expert guidance. IEEE TNNLS (2023)
86. Yang, X., Shih, S.M., Fu, Y., Zhao, X., Ji, S.: Your ViT is secretly a hybrid discriminative-generative diffusion model. arXiv:2208.07791 (2022)
87. You, H., Xiong, Y., Dai, X., Wu, B., Zhang, P., Fan, H., Vajda, P., Lin, Y.: Castling-vit: Compressing self-attention via switching towards linear-angular attention during vision transformer inference. In: IEEE CVPR (2023)
88. Zhang, T., Huang, H.Y., Feng, C., Cao, L.: Enlivening redundant heads in multi-head self-attention for machine translation. In: EMNLP (2021)
89. Zheng, H., Nie, W., Vahdat, A., Anandkumar, A.: Fast training of diffusion models with masked transformers. TMLR (2024)