FocusDiT: Masking Queries in Diffusion Transformers for Fine-grained Image Generation

Xueji Fang
Zhejiang University
Westlake University
fangxueji@zju.edu.cn

Liyuan Ma*
Westlake University
maliyuan@westlake.edu.cn

Jianhao Zeng
Westlake University
zengjianhao@westlake.edu.cn

Jinjin Cao
Zhejiang University
Westlake University
caojinjin@westlake.edu.cn

Mingyuan Zhou
Westlake University
zhoumingyuan@westlake.edu.cn

Guo-Jun Qi*
Westlake University
guojunq@gmail.com

Figure 1: Fine-grained text-to-image samples from our FocusDiT, showcasing its capabilities in attention to fine details and high image quality in various styles.

ABSTRACT

Diffusion transformer (DiT) has been widely adopted in the generative diffusion field, advancing the denoising of query tokens through attention and Feed-Forward (FFN) layers. FFN actually acts as the key-value vocabulary for decoding visual contents where the value embeds the visual semantical knowledge. We present that focusing on critical query tokens corresponding to more complex details and encouraging the model to improve these tokens is essential for fine-grained visual generation. To this end, we propose FocusDiT, which applies a Masking scheme to focus on critical query tokens that are exclusively fed into FFN. The masked queries can retrieve visual tokens from the FFN vocabularies, and use them to decode their visual details. Extensive text-to-image experiments validate the effectiveness of token masking in enhancing generative performance.

KEYWORDS

Diffusion Model, Transformer, Text-to-Image Generation, Token Selection

1 INTRODUCTION

Diffusion transformers [33] have become the state-of-the-art approach in visual generation, showing impressive performance in various tasks. An increasing amount of research has been devoted into the design of diffusion transformer architectures. For instance, DiG [48] introduces linear attention to address computational load issues, while Stable Diffusion 3 [15] enhances text-visual alignment

*Corresponding author.

by incorporating a multi-branch structure for text and visual interaction within the attention mechanism. Several studies also target the FFN, such as EC-DiT [40], which introduces Mixture-of-Experts (MoE) techniques into the DiT, scaling up the model by increasing the number of FFNs to improve performance. However, there is still a lack of comprehensive study and optimization of the FFN in DiT that plays a key role in decoding visual content.

In large transformer models, the FFN has been shown to function as a key-value vocabulary, and the DiT leverages the FFN to store visual tokens required for generating diverse visual semantics. These visual tokens, representing the visual semantical knowledge learned from the training data as presented in the left of Figure 2, are embedded into the FFN's weights and are subsequently accessed by the *query tokens* from preceding attention layers. We argue that critical query tokens characterizing fine-grained visual details should be selectively masked and fed into the FFN to retrieve more visual tokens for visual content decoding. However, the critical tokens are often overwhelmed by those corresponding to backgrounds or low-frequency regions of few visual details, wasting the full utilization of FFN's visual vocabulary to refine less important query tokens. This interference in utilization across FFN vocabulary hinders further performance improvement. Additionally, the FFN vocabulary is not fully leveraged as indicated in the right of Figure 2, with certain elements in shallow and deep layers seldom being utilized by the query tokens. This suggests that the decoding process in different FFN layers requires vocabularies of varying sizes.

To address these challenges, we propose **FocusDiT** which applies *Query Token Masking* to boost the FFN vocabulary utilization for critical tokens. The query mask distinguishes between tokens with complex structures and those with simpler ones, masking out the latter to avoid interference in decoding the visual content of critical tokens. Besides, we adjust the vocabulary size for each FFN layer to match the specific demands of the query token decoding process, while maintaining the total FFN parameter count. In particular, we allocate smaller vocabulary to shallow and deep layers handling simple content decoding and minimal details filling, reallocating the reduced capacity to the intermediate layers that decode both coarse and fine details. Further investigation shows that with accurate query mask selection, FocusDiT enhances fine-grained visual generation. Additionally, the query mask improves inference efficiency by skipping FFN layers where most mask values are close to zero. Both quantitative and qualitative results from text-to-image experiments show advantages over competing methods, confirming the effectiveness of our approach. Our contributions are summarized as follows:

- We propose a Query Token Masking strategy that identifies and prioritizes critical query tokens containing complex structures, enhancing vocabulary utilization for these tokens to capture fine-grained details. Besides, the query mask can be used to decrease inference cost, thereby enriching its practical significance.
- We propose a Vocabulary Redistribution (VR) scheme that reallocates vocabulary capacity across network layers, ensuring better alignment with the generation needs of query tokens, thereby enabling more efficient use of the total vocabulary.

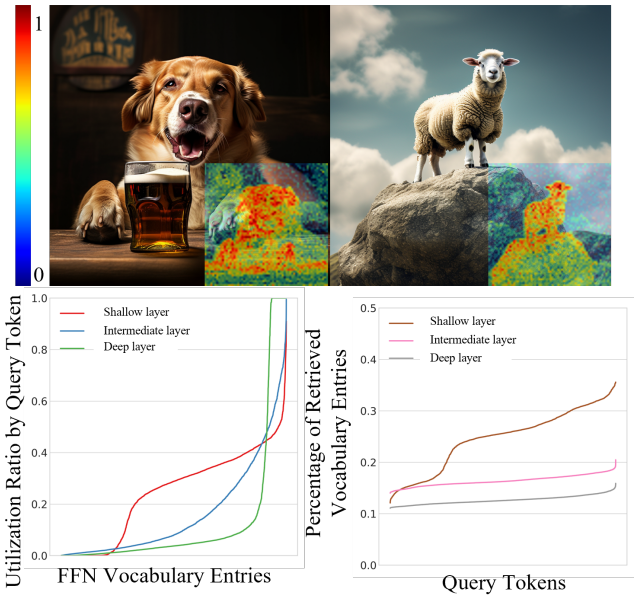

Figure 2: Top: Heatmaps indicating the number of entries from certain vocabulary subset retrieved by query tokens from different content, highlighting that certain vocabulary tokens correspond to specific semantic (e.g., animals). Bottom: Analysis of vocabulary utilization across layers and query tokens, based on 100 samples from DiT-based model [5].

- Extensive experiments and analyses on visual generation tasks demonstrate the superiority of FocusDiT in both visual quality and quantitative evaluation.

2 RELATED WORK

2.1 Generative Diffusion Models

Diffusion models [9, 22, 24, 28, 31, 39, 46] have shown impressive performance in generative tasks, trained with denoising objective $\mathcal{L}_{DM} = \mathbb{E}_{x_0,\theta,t}\left[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\right\|_2^2\right]$ with $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, timestep t, dataset sample x_0, timestep-related hyperparameter $\bar{\alpha}_t$, model parameters θ. DiT replaces the U-Net [37] backbone with transformer for a range of generative diffusion modeling tasks, yielding notable improvements in both performance and scalability. For instance, PixArt series [3–5] incorporates a cross-attention module for the text-to-image task, while Sora [2] further extends DiTs for text-to-video generation. Most improvements to DiTs have focused on optimizing the attention module [15, 26, 27, 48], with comparatively few studies exploring the enhancement of the FFN. Earlier approaches [16, 40] explored employing MoE to integrate multiple FFN expert modules, expanding the model without altering the internal structure of the FFN. However, these methods have not deeply analyzed how the internal mechanisms of the FFN affect the generative process, which is the focus of our work.

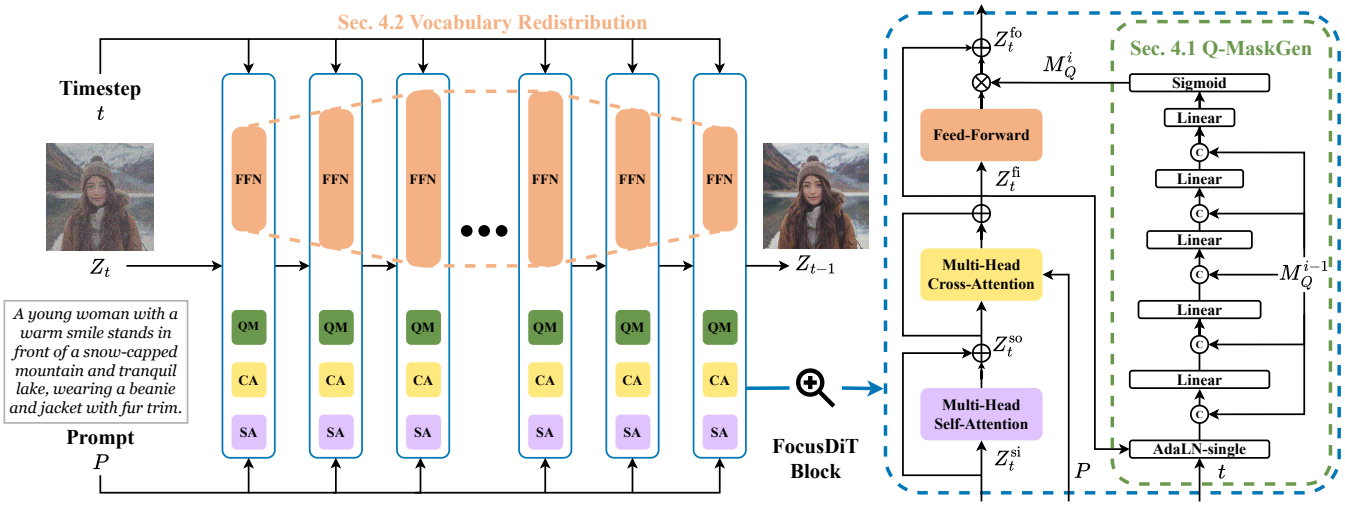

Figure 3: Framework of our proposed FocusDiT. Left: Overview of the main architecture, including multiple FocusDiT blocks. Right: Architecture of our query mask prediction network (Q-MaskGen).

2.2 FFN in Large Transformer Model

Research on large transformer model [11, 12, 17, 18, 23, 30, 43, 45] has shown that FFNs serve as the knowledge base for storing information from training dataset, retained in a key-value memory format during model training, enabling specific knowledge to be activated by input and mapped to the final generated output. Inspired by this, we examine the role of the FFN in DiT, which shares structural similarities to LLMs within the visual content generation process, finding that the FFN also acts as a vocabulary, storing various types of semantic concepts.

2.3 Token Selection

Token selection mechanisms are common adopted in generative visual and language modeling, enabling models to selectively distinguish and leverage tokens based on their importance [10, 20, 34, 42]. For instance, DynamicViT [36] selectively prunes less informative tokens to improve the model's efficiency. DiffRate [6] employs a token importance metric to identify and select the top-K tokens through token pruning and merging. DyDiT [47] dynamically selects well-denoised tokens and bypasses subsequent blocks, thereby reducing the model's computational load. These works mainly aim to improve the model's operational efficiency. In contrast, our query masking strategy directs the model to prioritize the vocabulary allocation to critical tokens, thereby improving visual generation quality.

3 PRELIMINARY

FFN in DiT actually serves as key-value vocabulary which stores the learned visual semantics [17, 18]. Specifically, FFN consists of two weight matrices $W_K \in \mathbb{R}^{d\times d_m}$, $W_V \in \mathbb{R}^{d_m\times d}$ and corresponding biases $b_K \in \mathbb{R}^{d_m}$, $b_V \in \mathbb{R}^{d}$. d_m actually represents the vocabulary size. The input query token $Z_t^{\text{fi}} \in \mathbb{R}^d$ interacts with the first weight matrix corresponding to *key* and then aggregates vocabulary entries saved in second *value* weight matrix W_V to produce final output Z_t^{fo}. FFN$(\cdot)$ calculation is denoted as:

$$\begin{aligned} Z_t^{\text{fo}} &= Z_t^{\text{fi}} + \text{FFN}(Z_t^{\text{fi}}) \\ &= Z_t^{\text{fi}} + [f(Z_t^{\text{fi}} \cdot W_K + b_K) \cdot W_V + b_V], \end{aligned} \tag{1}$$

where f is a non-linearity activation function such as GeLU.

Vocabulary utilization in FFN is insufficient and not distinguished for DiT. As shown in the bottom left of Figure 2, vocabulary utilization is insufficient, with many rarely activated during the generation process. Although DiT can access relevant visual tokens from the FFN for denoising, the allocation to query tokens shows minimal variation as shown in the bottom right of Figure 2, failing to distinguish between critical and non-critical tokens, which limits the ability of critical tokens to access more vocabulary entries. These limitations hinder the effective utilization of vocabulary and degrade generation quality.

4 METHODOLOGY

Based on the analyses above and with the aim of enhancing critical query token decoding, we formulate hypothesis that ***Can offering identified query tokens in FFN with a greater share of the vocabulary improve generation outcomes?*** To explore this, we design Query Mask (Section 4.1) to prioritize the critical query token generation and Vocabulary Redistribution (Section 4.2) to adjust the vocabulary size for more effective query token decoding.

4.1 Query Mask for Token Focus

The query mask is designed to guide critical tokens to query richer visual semantics from the FFN vocabulary to improve their decoded representation. This ensures that the model can focus on decoding query tokens for various visual details. To achieve this, the query mask dynamically predicts which tokens are responsible for complex details, marking them with a value close to 1.

The prediction of query mask considers the diffusion timestep since the query tokens need to focus on vary with different noise

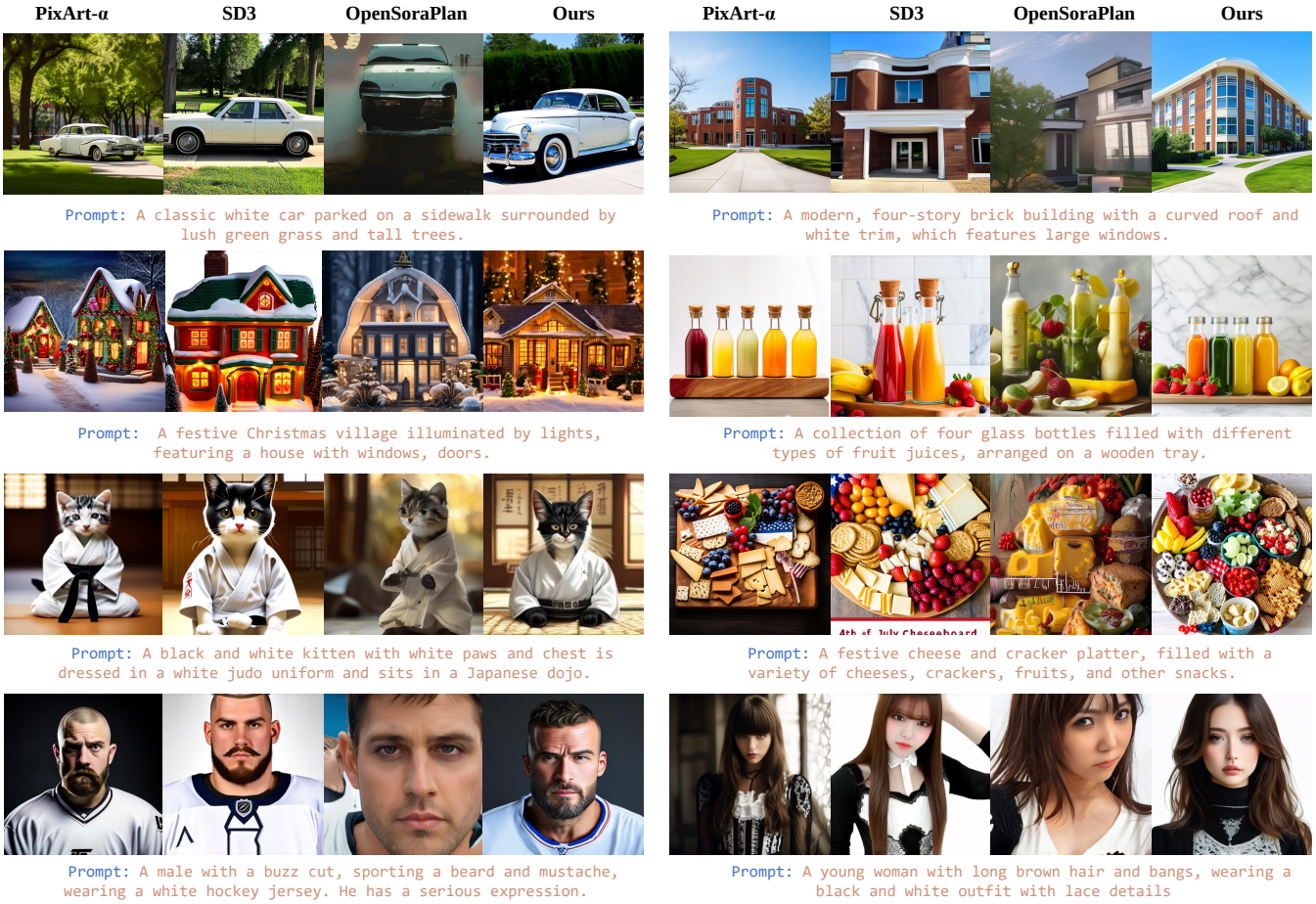

Figure 4: Qualitative comparison with PixArt-α, SD3 and OpenSoraPlan.

levels. At high noise levels, the information across query tokens is relatively uniform, so no special distinction is needed. However, as the noise level decreases, critical query tokens that represent complex details become harder to denoise, and should thus receive more attention compared to non-critical tokens. We thus explicitly incorporate the timestep into the query mask prediction using AdaLN-single [5], enabling the network to better adapt to the denoising stage. Additionally, given that query tokens are decoded sequentially across FFN blocks, we integrate the mask from the previous block into the query mask prediction of the current block. This helps maintain consistency in identifying critical tokens, facilitating more comprehensive denoising and improving the generation of complex details. Specifically, as illustrated in Figure 3, query mask prediction network (Q-MaskGen) takes the timestep t, cross-attention output Z_t^{fi} and previous query mask M_Q^{i-1} as the inputs, allowing prompt to better influence mask generation. Q-MaskGen is composed of several MLP layers and can be expressed as follows:

$$M_Q^i = \text{Q-MaskGen}(Z_t^{\text{fi}}, M_Q^{i-1}, t), i = 1, ..., N, M_q^0 = I, \tag{2}$$

where the N denotes the number of FocusDiT blocks.

The query mask enables the model to exclude non-critical query tokens in the current timestep and block, ensuring that only critical tokens make full use of the FFN to decode visual details. In particular, as illustrated in Figure 3, the input features Z_t^{fi} are passed through the FFN to output updates, which are adaptively adjusted by the query mask M_Q and added to the residual to produce the final result Z_t^{fo}. The calculation is as follows:

$$\begin{aligned} Z_t^{\text{fo}} &= Z_t^{\text{fi}} + M_Q \cdot \text{FFN}(Z_t^{\text{fi}}) \\ &= Z_t^{\text{fi}} + M_Q \cdot [f(Z_t^{\text{fi}} \cdot W_K + b_K) \cdot W_V + b_V]. \end{aligned} \tag{3}$$

In fact, the query mask reweighs the influence of critical tokens on FFN vocabulary weight updates during backpropagation, preserving gradients from critical tokens to encourage the encoding of visual tokens corresponding to more complex visual details in the FFN vocabulary. The backpropagation is conducted as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}_{DM}}{\partial W_K} &= M_Q \cdot \frac{\partial \mathcal{L}}{\partial Z_t^{\text{fo}}} \cdot \frac{\partial \text{FFN}(Z_t^{\text{fi}})}{\partial W_K}, \\ \frac{\partial \mathcal{L}_{DM}}{\partial W_V} &= M_Q \cdot \frac{\partial \mathcal{L}}{\partial Z_t^{\text{fo}}} \frac{\partial \cdot \text{FFN}(Z_t^{\text{fi}})}{\partial W_V}. \end{aligned} \tag{4}$$

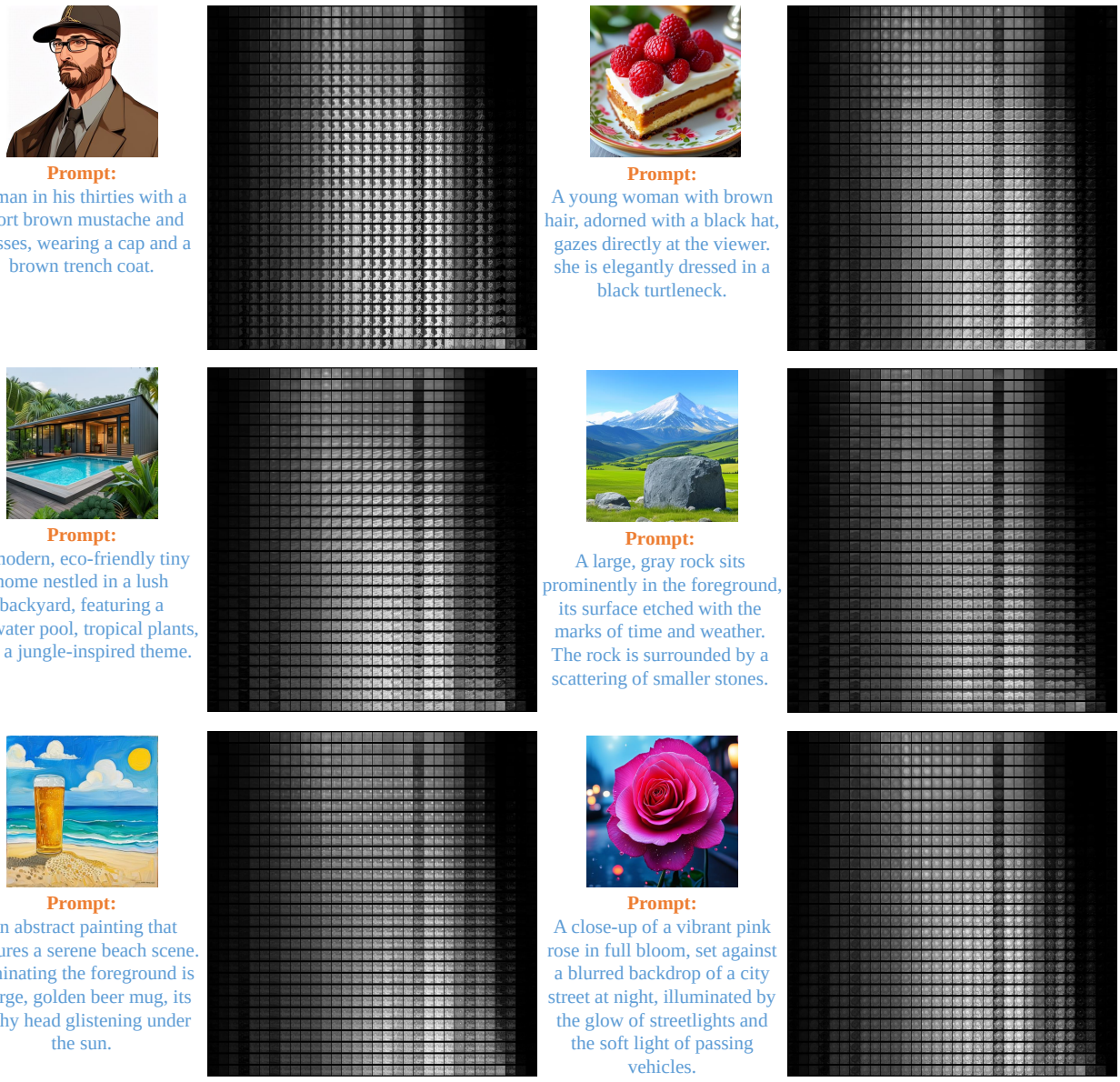

Figure 5: Visualization of query masks across different FocusDiT blocks and denoising timesteps. From left to right, the grid represents the progression from the first Transformer block to the last, while from top to bottom, it corresponds to increasing denoising timesteps. Lighter regions highlight critical query tokens that play a more significant role in decoding, whereas darker regions indicate less influential tokens.

4.2 Vocabulary Redistribution

The FFN in DiT serves as vocabulary which provides query tokens with varying amounts of visual semantics embedded in visual tokens during generation. Therefore, designing a vocabulary size that aligns with token generation requirements can benefit the generation process. The FFN vocabularies in shallow and deep layers require fewer capacity, as query tokens in these layers only generate simple content or complement a small amount of visual details. In contrast, the middle layers, which handle both coarse and fine-grained structure generation, demand a larger vocabulary to accommodate a wider range of semantic information.

To this end, redistributing the vocabulary capacity across different layers is reasonable. As shown in Figure 3, we design a vocabulary redistribution strategy that reduces the vocabulary size of shallow and deep layers of FocusDiT, which corresponds to the d_m in W_V of FFN. Then the reduced FFN capacity from the shallow and deep layers are reallocated to the FFN vocabulary of middle-layers, while maintaining the overall model parameter count.

Specifically, the i-th FFN in our FocusDiT is defined as follows:

$$\begin{cases} \{W_{K_i}^{M\to P+(i-1)\times\Delta}, W_{V_i}^{P+(i-1)\times\Delta\to M}\}, & i \le L/2, \\ \{W_{K_i}^{M\to P+(L-i)\times\Delta}, W_{V_i}^{P+(L-i)\times\Delta\to M}\}, & i > L/2. \end{cases} \tag{5}$$

where P is the internal dimension of the first FFN, and Δ denotes a constant change in size between adjacent FFNs.

To ensure that the parameters count O^* of our FFNs after vocabulary redistribution remains consistent with the baseline, we enforce the restriction:

$$O^* = 2\sum_{i=1}^{L/2} 2\big(M \times (P + (i-1)\Delta)\big) = 2LMN. \tag{6}$$

In the baseline configuration, $N = 4M$. We set $P = 3M$, then we can derive $\Delta = \lfloor \frac{2}{(L/2-1)} M \rfloor$.

This transforms the vocabulary size distribution from an even spread to a spindle shape and optimizes capacity for different FFN layers.

5 EXPERIMENT

5.1 Implementation Details

We present the details of our method, including datasets, model architecture, training and evaluation processes.

Datasets. Our model is trained on both real-world and synthesized text-image pairs. The real-world images are sourced from three datasets: Coyo-HD-11M [13], LAION-Aesthetics-V1-120M [38] and LAION-Art-8M [38], all chosen for their high aesthetic quality. The synthesized images come from Midjourney-Niji-1M [14], generated using Midjourney-V6. Additionally, we expand the dataset with an internal collection of 500K text-image pairs.

Training Details. We use Flan-T5-XXL [8] for prompt encoding and VAE [41] for image encoding, applying 8×8 spatial downsampling while maintaining reconstruction quality. The model is first trained from scratch on 256×256 resolution with 2048 batch size for diverse concepts learning, then on 512 × 512 with 1280 batch size for high-quality improvement. The training process takes around 60k steps, utilizing the AdamW [29] optimizer and a learning rate of 2×10^{-4}.

Evaluation Metrics. We assess the generation quality using four established metrics: Fréchet Inception Distance (FID)[21] measures the similarity between the feature distributions of generated and real images using a pre-trained Inception network, where a lower FID indicates a closer match to real images. CLIPScore[32] evaluates the semantic alignment between generated images and their corresponding prompts using the CLIP model [35], with higher scores indicating better alignment. GenEval[19] assesses the compositional quality of generated images, including object co-occurrence, spatial arrangement, object count, and color accuracy. ImageReward[44] provides a perceptual evaluation of image quality based on human preference signals, capturing aspects such as aesthetics and coherence. In addition to these metrics, we introduce Structural Clarity and Textural Fidelity (SCTF) to assess the perceptual quality of generated images. SCTF is measured using Vision-Language Models (VLMs) such as Qwen2.5-VL[1] and InternVL-2.5[7], which assign a score from 0 to 10 based on the clarity of structural elements and the fidelity of textures, ensuring the absence of artifacts or hallucinations. Specially, for each model and configuration, we compute the **SCTF score** as the **mean score** across all generated images:

$$SCTF = \frac{1}{N}\sum_{i=1}^{N} s_i \tag{7}$$

where N is the total number of images, and s_i is the SCTF score assigned to the i-th image by the VLM. To quantify the variability of SCTF scores, we report the **standard error of the mean (SEM)**, which is computed as:

$$SEM = \frac{\sigma}{\sqrt{N}} \tag{8}$$

where σ is the standard deviation of the SCTF scores across all images. The SEM provides an estimate of the uncertainty in the mean SCTF score, allowing for more reliable comparisons between different model configurations.In all experiments, we generated a total of 5,120 images for evaluation, i.e., $N = 5120$. SCTF provides a fine-grained evaluation of image quality beyond traditional statistical measures.

5.2 Qualitative Analysis

In comparison to other DiT-based models such as PixArt-α, SD3, and OpenSoraPlan [25], our proposed FocusDiT demonstrates the superiority in various aspects of image generation as shown in Figure 4. FocusDiT excels in generating finer details, especially in character features like hair textures, which are more intricate and realistic compared to the other models. It also produces more structurally sound and coherent architectural designs, maintaining better proportions and plausibility in generated buildings. Additionally, when tasked with generating non-realistic or imaginative scenes, such as "a kitten dressed in judo uniform," FocusDiT delivers results with superior structural refinement and more nuanced details, effectively capturing the complexity of abstract concepts. These results underline the strengths of FocusDiT in both realistic and creative image generation, offering enhanced fidelity in character details and structural accuracy in complex environments.

Query Masks Visualization. As shown in Figure 5, we visualize the Query Masks predicted by our Q-MaskGen across each block through 30 denoising timesteps. From the perspective in the FocusDiT blocks, the query masks are predicted from the first block to the last sequentially, where we observe that both shallow and deep blocks tend to mask a significant portion of the queries representing high-frequency details such as car edges. Specifically, in the shallowest and deepest blocks, the query mask effectively masks out almost all query tokens during the FFN update, indicating that decoding these tokens does not rely on these FFNs. Based on this observation, we are able to accelerate the inference process by skipping layers with query masks close to zero, as discussed in section 6. From a the denoising timestep perspective, since the query tokens gradual transit from a gaussian noise to a structured image, as the denoising progresses, the contours of the foreground and background become increasingly refined. Notably, we also observe that the selected critical token does not continuously focus on the generation of fine details in the subject at different timesteps. Timesteps corresponding to high noise level are more focused on generating

Table 1: Quantitative comparisons of text-to-image generation. Compared to PixArt-α, our model uses less than 20% of its computational resources, yet outperforms it in terms of FID and GenEval benchmark, and achieves comparable performance on CLIPScore.

Model	FID↓	CLIPScore↑	ImageReward↑	GenEval↑	Training Time	Training Iterations
Opensora-Plan	62.41	0.217	-1.85	0.02	–	346k
PixArt-α	30.50	0.315	0.74	0.48	753 A100 days	–
SD3	23.15	0.321	0.92	0.70	–	$\geq$ 500k
FocusDiT (Ours)	27.81	0.307	0.29	0.57	**156 A100 days**	**72k**

the background details, while timesteps with lower noise shift their attention toward the foreground and finer structures, as evidenced in Figure 5.

Vocabulary Utilization. As shown in Figure 6, we visualize and compare the utilization rates of query tokens at different layers between FocusDiT and DiT. In shallow blocks, FocusDiT demonstrates a significantly higher utilization rate compared to DiT, which can be attributed to our approach of redistributing vocabulary capacity across the layers. For the deep layers, FocusDiT demonstrates a relatively higher utilization rate. This redistribution allows FocusDiT to make more efficient use of the entire vocabulary.

5.3 Quantitative Evaluation

From the quantitative analysis in Table 1, it is observed that although the foundational architecture of FocusDiT is quite similar to that of OpenSoraPlan and PixArt-alpha, our model significantly outperforms OpenSoraPlan and even surpasses PixArt-alpha. Notably, this achievement comes despite PixArt-alpha being trained on a large amount of internal data, whereas FocusDiT primarily relies on open-source datasets. However, the limited scale and diversity of available open-source data have prevented our model from surpassing SD3. We believe that with access to larger and more diverse datasets and additional training time, FocusDiT's performance could be further enhanced to close the remaining gap with SD3 in future work.

User Study. We conduct a user study for evaluating fine-grained structure in image generation, which compares the performance of our image generation method with two competing models: PixArt-α [5] and SD3 [15]. After 660 times comparison made by users, the results presented in Figure 7 clearly demonstrate that our proposed method outperforms both PixArt-α and SD3 in human preference quality. Specifically, compared to PixArt-α, our method achieves a 52.6% win rate, and when compared to SD3, our method surpasses it with a 54.5% win rate. This suggests that our image generation method not only matches but also outperforms the other models for less artifacts and better fine-grained details with the given prompts.

5.4 Ablation Study

To prevent unnecessary overhead due to the increase in model parameters, we adopted a simple Multi-Layer Perceptron (MLP) structure for MaskGen. We conducted carefully designed ablation experiments to demonstrate the effectiveness of our architecture. First, we assessed the impact of the number of MLP layers by testing 2-layer, 5-layer, and 10-layer configurations, with the 2-layer and 10-layer networks representing shallower and deeper designs, respectively. The 5-layer MLP achieved the best balance between model complexity and performance, as shown in Table 2. Next, adding a Dropout layer encouraged the model to mask out more queries during training, enhancing focus on critical queries and refining the mask generation process. Additionally, we also introduced a prior mask from the previous layer into MaskGen. This enhancement allowed the model to better concentrate on previously masked regions, resulting in higher-quality mask generation and improved final image generation. While the FFN in our VR structure showed a slight degradation in image quality when trained for the same 30k steps, as illustrated in Figure 6, the model's performance gradually improved with continued training. This improvement was attributed to better vocabulary utilization in the FFN, which ultimately enhanced the final generation performance. Despite a minor drop in the FID metric, our VR-enhanced architecture outperformed the non-VR structure in terms of Structural Clarity and Texture Fidelity (SCTF), demonstrating that incorporating VR contributes to clearer structural details and more plausible textures in the generated images. Through these ablation experiments, we demonstrate that each design choice in our MaskGen module contributes significantly to the overall performance of FocusDiT, ensuring both efficient and high-quality image generation.

6 INFERENCE ACCELERATION

The query mask reflects how much a token relies on the FFN's vocabulary. We believe that the lower the mean of the query mask, the smaller the contribution of the current FFN to token denoising, therefore the model can skip these FFNs to achieve accelerated inference while preserving its overall performance. We analyze the tradeoff between the model's overall performance and inference efficiency by setting a threshold. As shown in Figure 8, when the threshold is low, fewer FFNs are skipped, and the model's overall performance is preserved as much as possible; when the threshold is high, more FFNs are skipped, which significantly improves the model's inference efficiency, but severely impacts its overall performance. Besides, We observe that lower thresholds have little impact on network performance, but once the threshold exceeds a certain range, the model's performance drops significantly. With a threshold of 0.15, the time to generate 16 samples decreases from 58.69s to 50.66s without significant performance drop.

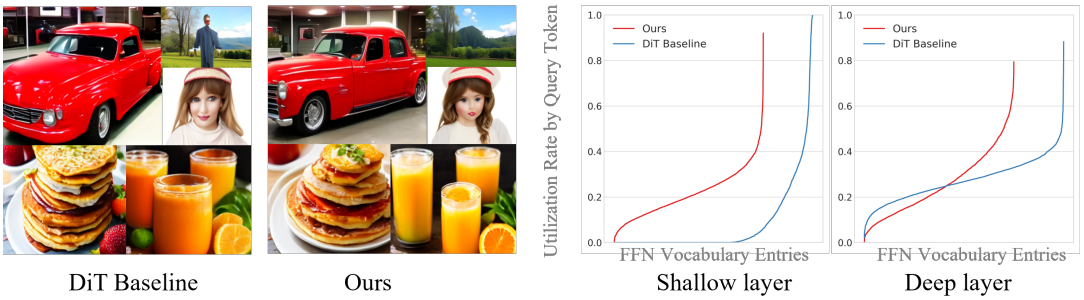

Figure 6: Comparison of vocabulary utilization between the DiT baseline and FocusDiT across various prompts.

Table 2: Model variants of FocusDiT at 30k training steps. Full Model denotes B_2 + Dropout + Premask + VR.

	Models	Image Quality			Structural Clarity and Textural Fidelity (SCTF)↑			
		FID↓	CLIPScore↑	ImageReward↑	InternVL2.5-4B	InternVL2.5-8B	Qwen2.5-VL-3B	Qwen2.5-VL-7B
A	DiT Baseline	31.29	0.310	-0.71	7.5959±0.0208	8.1184±0.0119	6.4705±0.0193	8.1785±0.0126
B_1	A + 2layers	33.59	0.306	-0.87	7.4889±0.0224	8.0793±0.0129	6.5273±0.0198	8.1488±0.0134
B_2	A + 5layers	31.42	0.314	-0.66	7.7328±0.0195	8.2113±0.0106	6.5854±0.0201	8.2137±0.0127
B_3	A + 10layers	31.58	0.312	-0.65	7.7705±0.0172	8.2203±0.0099	6.5744±0.0187	8.2480±0.0103
C	B_2 + Dropout	31.00	0.311	-0.70	7.6693±0.0195	8.1773±0.0111	6.5008±0.0197	8.1625±0.0124
D	B_2 + Premask	32.19	0.315	-0.57	7.7668±0.0175	8.2256±0.0098	6.6541±0.0181	8.2824±0.0110
E	B_2 + VR	34.02	0.310	-0.76	7.7625±0.0186	8.2172±0.0106	6.5779±0.0191	8.2621±0.0115
F	C + D + E	32.93	0.315	-0.62	7.7922±0.0171	8.2602±0.0091	6.6121±0.0182	8.2537±0.0101
G	C + D + VR-shared-C	32.93	0.315	-0.62	7.7922±0.0171	8.2602±0.0091	6.6121±0.0182	8.2537±0.0101
H	C + D + VR-shared-2C	32.93	0.315	-0.62	7.7922±0.0171	8.2602±0.0091	6.6121±0.0182	8.2537±0.0101
I	C + D + VR-shared-3C	32.93	0.315	-0.62	7.7922±0.0171	8.2602±0.0091	6.6121±0.0182	8.2537±0.0101

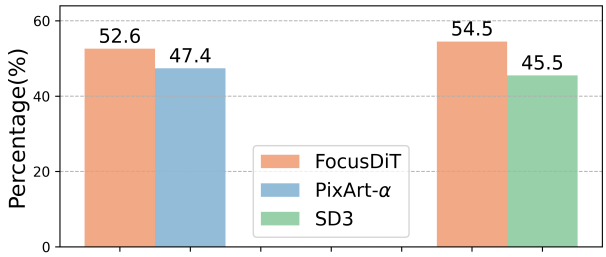

Figure 7: User Study. Our FocusDiT outperforms both PixArt-α and SD3 in human preference evaluation.

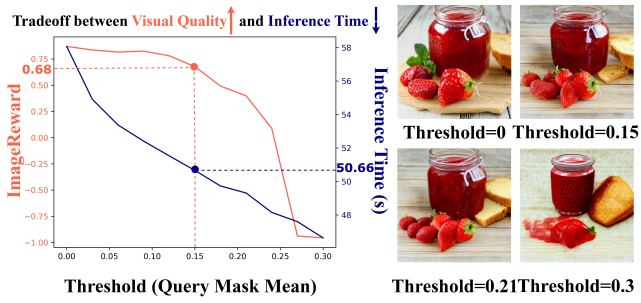

Figure 8: Inference time reduction by skipping FFN calculation guided by the mean of query masks.

Parameter Reduction. To further explore the impact of the less important FFN on token denoising, as shown in Figure 9, we calculate the average of the mean values of query masks across 200 different prompts, at various timesteps and different layers of FocusDiT. We find that the mean value of the query mask remains relatively consistent across different timesteps. However, there is a notable variation in this value across different layers of FocusDiT. Based on this observation, we instruct the model to skip specific FFNs across all timesteps, in other words, we prune the model's parameter by removing some of the FFN parameters in certain layers. As shown in Figure 9, as the number of deleted layers of FocusDiT increases, the overall performance of the model initially shows no significant change, then suddenly decreases substantially. After removing 8 relatively unimportant FFNs, the model's parameter count is reduced to 88% of the original, while the overall performance of the model shows no significant degradation.

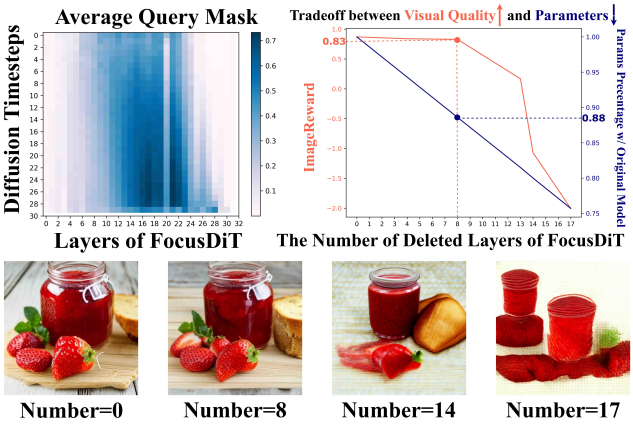

Figure 9: Parameters reduction by removing FFN guided by the mean of query mask.

7 CONCLUSION

In this paper, we introduce a Query Token Masking strategy aimed at enhancing the utilization of critical query tokens with complex structures, allowing the model to capture finer details in visual generation tasks. Additionally, we propose a Vocabulary Redistribution (VR) scheme that reallocates vocabulary capacity across network layers, ensuring better alignment with the specific needs of these query tokens. This approach not only improves vocabulary efficiency but also reduces inference costs, ultimately leading to more accurate and computationally efficient models for image generation, particularly in fine-grained details. Furthermore, we also introduce a threshold-based method that skips certain Feed-Forward Networks based on the query mask, balancing inference speed and performance.

REFERENCES

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators. (2024). https://openai.com/research/video-generation-models-as-world-simulators

[3] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692* (2024).

[4] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. 2024. Pixart-δ: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252* (2024).

[5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. 2023. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426* (2023).

[6] Mengzhao Chen, Wenqi Shao, Peng Xu, Mingbao Lin, Kaipeng Zhang, Fei Chao, Rongrong Ji, Yu Qiao, and Ping Luo. 2023. Diffrate: Differentiable compression rate for efficient vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17164–17174.

[7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271* (2024).

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. https://doi.org/10.48550/ARXIV.2210.11416

[9] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 9 (2023), 10850–10869.

[10] Yutao Cui, Tianhui Song, Gangshan Wu, and Limin Wang. 2024. Mixformerv2: Efficient fully transformer tracking. *Advances in Neural Information Processing Systems* 36 (2024).

[11] Jeff Da, Ronan Le Bras, Ximing Lu, Yejin Choi, and Antoine Bosselut. 2021. Analyzing commonsense emergence in few-shot knowledge models. *arXiv preprint arXiv:2101.00297* (2021).

[12] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696* (2021).

[13] Caption Emporium. 2024. coyo-hd-11m-llavanext. https://huggingface.co/datasets/CaptionEmporium/coyo-hd-11m-llavanext.

[14] Caption Emporium. 2024. midjourney-niji-1m-llavanext. https://huggingface.co/datasets/CaptionEmporium/conceptual-captions-cc12m-llavanext.

[15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

[16] Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. 2024. Scaling diffusion transformers to 16 billion parameters. *arXiv preprint arXiv:2407.11633* (2024).

[17] Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680* (2022).

[18] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913* (2020).

[19] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2024. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems* 36 (2024).

[20] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. 2021. Transformer in transformer. *Advances in neural information processing systems* 34 (2021), 15908–15919.

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[23] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics* 8 (2020), 423–438.
[24] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems* 35 (2022), 26565–26577.
[25] PKU-Yuan Lab and Tuzhan AI etc. 2024. *Open-Sora-Plan.* https://doi.org/10.5281/zenodo.10948109
[26] Black Forest Labs. 2024. *FLUX.1-dev Model Documentation.* https://huggingface.co/black-forest-labs/FLUX.1-dev Accessed: 2024-11-09.
[27] Black Forest Labs. 2024. *FLUX.1-schnell Model Documentation.* https://huggingface.co/black-forest-labs/FLUX.1-schnell Accessed: 2024-11-09.
[28] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022).
[29] I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
[30] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems* 35 (2022), 17359–17372.
[31] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*. PMLR, 8162–8171.
[32] I. Pavlov, A. Ivanov, and S. Stafievskiy. 2023. Text-to-Image Benchmark: A benchmark for generative models. https://github.com/boomb0om/text2image-benchmark. Version 0.1.0.
[33] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4195–4205.
[34] Jonathan Pilault, Mahan Fathi, Orhan Firat, Chris Pal, Pierre-Luc Bacon, and Ross Goroshin. 2024. Block-state transformers. *Advances in Neural Information Processing Systems* 36 (2024).
[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
[36] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems* 34 (2021), 13937–13949.
[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. 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*. Springer, 234–241.
[38] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.
[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).
[40] Haotian Sun, Bowen Zhang, Yanghao Li, Haoshuo Huang, Tao Lei, Ruoming Pang, Bo Dai, and Nan Du. 2024. EC-DIT: Scaling Diffusion Transformers with Adaptive Expert-Choice Routing. *arXiv preprint arXiv:2410.02098* (2024).
[41] Genmo Team. 2024. Mochi 1. https://github.com/genmoai/models.
[42] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).
[43] Jonas Wallat, Jaspreet Singh, and Avishek Anand. 2021. BERTnesia: Investigating the capture and forgetting of knowledge in BERT. *arXiv preprint arXiv:2106.02902* (2021).
[44] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2024. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2024).
[45] Yunzhi Yao, Shaohan Huang, Li Dong, Furu Wei, Huajun Chen, and Ningyu Zhang. 2022. Kformer: Knowledge injection in transformer feed-forward layers. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 131–143.
[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.
[47] Wangbo Zhao, Yizeng Han, Jiasheng Tang, Kai Wang, Yibing Song, Gao Huang, Fan Wang, and Yang You. 2024. Dynamic diffusion transformer. *arXiv preprint arXiv:2410.03456* (2024).
[48] Lianghui Zhu, Zilong Huang, Bencheng Liao, Jun Hao Liew, Hanshu Yan, Jiashi Feng, and Xinggang Wang. 2024. DiG: Scalable and Efficient Diffusion Models with Gated Linear Attention. *arXiv preprint arXiv:2405.18428* (2024).