Diff-MoE: Diffusion Transformer with Time-Aware and Space-Adaptive Experts

Kun Cheng [*1] **Xiao He** [*1] **Lei Yu** [2] **Zhijun Tu** [2] **Mingrui Zhu** [1] **Nannan Wang** [†1] **Xinbo Gao** [1] **Jie Hu** [2]

Abstract

Diffusion models have transformed generative modeling but suffer from scalability limitations due to computational overhead and inflexible architectures that process all generative stages and tokens uniformly. In this work, we introduce Diff-MoE, a novel framework that combines Diffusion Transformers with Mixture-of-Experts to exploit both temporarily adaptability and spatial flexibility. Our design incorporates expert-specific timestep conditioning, allowing each expert to process different spatial tokens while adapting to the generative stage, to dynamically allocate resources based on both the temporal and spatial characteristics of the generative task. Additionally, we propose a globally-aware feature recalibration mechanism that amplifies the representational capacity of expert modules by dynamically adjusting feature contributions based on input relevance. Extensive experiments on image generation benchmarks demonstrate that Diff-MoE significantly outperforms state-of-the-art methods. Our work demonstrates the potential of integrating diffusion models with expert-based designs, offering a scalable and effective framework for advanced generative modeling. The code is available at https://github.com/kunncheng/Diff-MoE.

1. Introduction

Diffusion models (Ho et al., 2020; Song et al., 2020) have revolutionized generative modeling, achieving state-of-the-art results across various tasks, including image generation (Dhariwal & Nichol, 2021; Rombach et al., 2022; Chen et al., 2024; Esser et al., 2024), video generation (Blattmann et al., 2023; Yang et al., 2024b) and editing (Brooks et al., 2023; Qi et al., 2023). However, their scalability remains limited by significant computational costs. Despite recent advancements, such as Diffusion Transformers (DiT) (Peebles & Xie, 2023), the representation capacity and scalability of these models is constrained by two inherent limitations. First, the densely activated nature of DiT requires every parameter to be utilized across all stages of generation, regardless of their relevance to the specific task or input, resulting in excessive computational overhead. Second, DiT employs a monolithic design that processes all diffusion timesteps uniformly, assuming that the same network components can handle both the coarse-grained global structures of early diffusion stages and the fine-grained local refinements in later stages. This lack of timestep-specific adaptation prevents the model from specializing in the distinct requirements of different generative phases, ultimately hindering its representational capacity.

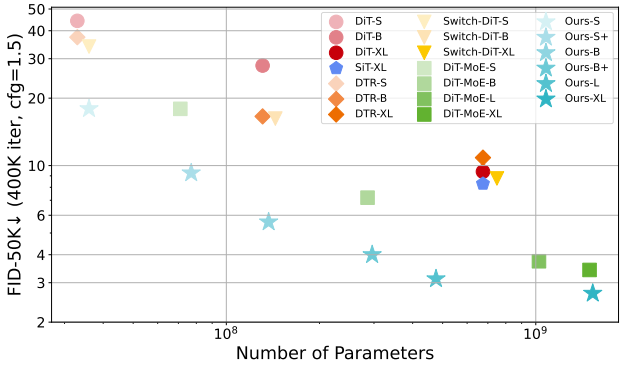


Figure 1. **Comparison of FID performance and Parameters across different sizes.** We compare the class of DiT models trained for 400K iterations on ImageNet 256×256 generation with $cfg = 1.5$. Diff-MoE achieves better FID across different sizes.

Mixture-of-Experts (MoE) architectures have emerged as a scalable paradigm for deep learning (Shazeer et al., 2017; Dai et al., 2024), offering computational efficiency through dynamic parameter activation. By assigning distinct "experts" to process different parts of the input, MoE models unlock unprecedented capacity without proportionally increasing computational cost. While MoE models excel in language tasks by routing inputs to specialized sub-networks ("experts"), their potential in diffusion-based generative

[*]Equal contribution [1]State Key Laboratory of Integrated Services Networks, Xidian University [2]Huawei Noah's Ark Lab. Correspondence to: Nannan Wang <nnwang@xidian.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

frameworks remains underexplored. Prior works either partition experts across denoising stages (temporal) (Balaji et al., 2022; Feng et al., 2023) or spatial tokens (Fei et al., 2024; Sun et al., 2024), but none efficiently integrate both dimensions, leaving untapped performance gains, as illustrated in Figure 2.

In this paper, we propose Diff-MoE, a diffusion framework that integrates timestep-aware and space-adaptive experts within Diffusion Transformers. Unlike naive multi-expert fusion methods (Xue et al., 2024) or rigid DiT-MoE hybrids (Fei et al., 2024), our approach introduces expert-specific timestep conditioning: each spatial expert dynamically adapts to the current denoising stage via lightweight embeddings, enabling specialized processing of tokens based on both spatial content (e.g., object edges vs. textures) and temporal context (early-stage global structures vs. late-stage refinements).

Furthermore, existing MoE designs often lack global input awareness, largely due to their application on autoregressive generation paradigms commonly used in language tasks. We propose an adaptive feature recalibration mechanism tailored specifically for image generation tasks, enhancing the representational capacity of expert modules. By incorporating globally contextual embeddings into the high-dimensional spaces within experts, this mechanism dynamically adjusts the contribution of different feature channels based on their relevance to the input distribution. This recalibration strengthens the channel mixers' ability to model complex dependencies, emphasize informative features, and suppress noise, ultimately improving generation quality. To further mitigate parameter overhead, we project the expert-specific timestep conditioning and global context extraction processes into a compact, low-rank subspace, ensuring efficient scaling without sacrificing performance.

Our framework achieves significant improvements over both dense and expert-based sparse DiTs. Extensive experiments on standard benchmarks validate the superiority of Diff-MoE, achieving state-of-the-art results in quantitative metrics. Our contributions can be summarized as follows:

- A unified MoE-diffusion architecture optimizing temporal adaptation and spatial specialization through expert-specific timestep conditioning. This strategy enables dynamic expert adaptation across different stages of the diffusion process while keeping token-specific flexiblity.
- We introduce a global feature recalibration mechanism, dynamically adjusting the contributions of feature channels based on their input relevance, empowering channel mixers to model complex dependencies.
- To reduce parameter overhead, we employ a low-rank technique that enables compact and efficient modulation without sacrificing expressiveness, ensuring scalability for large-scale models. Extensive experiments show that Diff-MoE significantly outperforms both dense DiT and state-of-the-art expert-based methods at the same parameter scale on conditional image generation tasks.

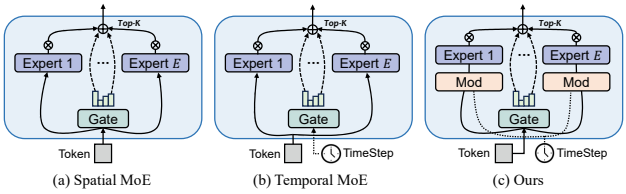


Figure 2. **The comparison from mainstream diffusion-based MoE methods to the proposed Diff-MoE.** (a) Standard spatial MoE: Experts are assigned based solely on spatial locations. (b) Temporal MoE: Experts are selected using a Top-K mechanism based on timestep information, allowing the model to focus on the dynamics of the diffusion process at different stages. (c) The proposed Diff-MoE: Combines both spatial content and timestep information, modulating the input token to enable specialized processing of tokens based on both spatial content and timestep.

2. Related Work

2.1. Diffusion Transformer

Diffusion models (Ho et al., 2020; Rombach et al., 2022) have surpassed Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) as the dominant paradigm for generative modeling, owing to their ability to capture complex data distributions and stable training dynamics. These models achieve state-of-the-art results across diverse domains, including image (Rombach et al., 2022; Chen et al., 2024), video (Mei & Patel, 2023), and 3D generation (Poole et al., 2022; Wang et al., 2024). Early architectures predominantly employed U-Net backbones (Ronneberger et al., 2015; Ho et al., 2020; Nichol & Dhariwal, 2021), but recent advances like Diffusion Transformers (DiT) (Peebles & Xie, 2023) have demonstrated superior scalability by integrating transformer designs. Building on DiT, Stable Diffusion 3 (SD3) (Esser et al., 2024) introduced a Multi-modal Diffusion Transformer (MM-DiT), scaling to 8B parameters for text-to-image synthesis. FLUX further expanded this to 12B parameters (Labs, 2023), showcasing DiT's potential for high-resolution generation. However, scaling these dense architectures—where all parameters are activated per sample and per timestep (Hatamizadeh et al., 2024; Liu et al., 2024b; Karras et al., 2022)—incurs prohibitive computational costs, limiting practical deployment.

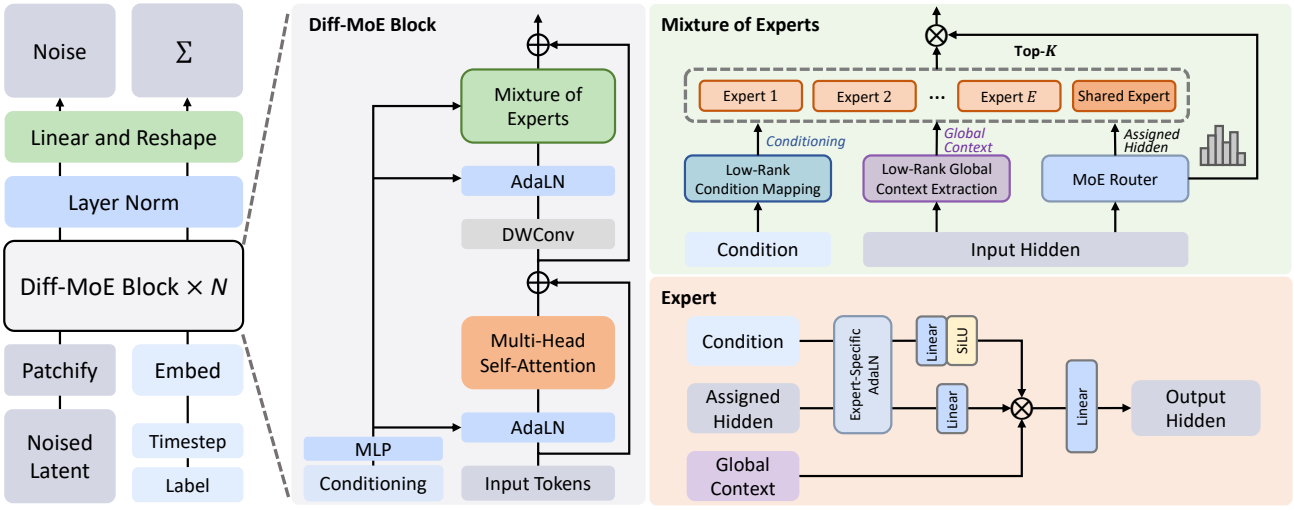


Figure 3. **Overview of the Diff-MoE architecture.** Left: The Diff-MoE architecture extends the Diffusion Transformer (DiT) framework (Peebles & Xie, 2023). Inputs are first patchified into latent tokens, then processed through a series of transformer blocks. Right: We replace the dense MLP in the original DiT with a carefully designed mixture of MLPs, enabling temporal adaptability (experts adjust to diffusion timesteps) and spatial flexibility (token-specific expert routing).

2.2. Mixture of Experts

Mixture-of-Experts (MoE) architectures (Shazeer et al., 2017; Dai et al., 2024) enhance model capacity by dynamically activating specialized sub-modules for distinct inputs, balancing expressiveness and efficiency. In the natural language processing (NLP) field, several studies have enhanced the core component of MoE, including the optimization of routing mechanisms (Zhou et al., 2022; Huang et al., 2024) and the architectural design of expert models (Dai et al., 2024; Zoph et al., 2022). These improvements have successfully enabled large language models (LLM) (Dai et al., 2024; Yang et al., 2024a) to expand their capacity while minimizing computational overhead. Inspired by the successful application of MoE in NLP, the computer vision community has begun adopting MoE to enhance the capacity of generative models, particularly diffusion models, which show significant potential. Time-centric approaches like DTR (Park et al., 2023) treat denoising as multi-task learning through channel masking, while Switch-DiT (Park et al., 2024) employs sparse temporal gating, and MEME (Lee et al., 2024) assigns experts to timestep ranges. While these partially enhance capacity, their exclusive focus on temporal partitioning ignores spatial complexity. Spatial MoE variants such as DiT-MoE (Fei et al., 2024) (16.5B parameters) and EC-DiT (Sun et al., 2024) (97B parameters via expert-choice routing) address token-level diversity but fail to adapt to temporal dynamics of the denoising process. Our approach bridges this divide through joint spatiotemporal expert coordination, demonstrating efficient scalability.

3. Preliminaries

3.1. Diffusion Models

Inspired by non-equilibrium thermodynamics, diffusion models learn to generate data by gradually reversing a stochastic noise process. In the forward process, data samples $\boldsymbol{x}_0$ from the target distribution $p(\boldsymbol{x}_0)$ are progressively corrupted by Gaussian noise through a series of timesteps t. The resulting noisy samples $\boldsymbol{x}_t$ are computed as follows:

$$q\left(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}\right)=\mathcal{N}\left(\boldsymbol{x}_t;\sqrt{\alpha_t}\boldsymbol{x}_0,\left(1-\alpha_t\right)\boldsymbol{I}\right), \quad (1)$$

where α_t defines the noise schedule. After a sufficient number of timesteps, the data distribution approaches a standard Gaussian.

The denoising process aims to reverse the forward diffusion by learning a parameterized noise predictor $\epsilon_\theta(\boldsymbol{x}_t,t)$. This predictor estimates the noise added at each step, enabling the recovery of $\boldsymbol{x}_0$ from $\boldsymbol{x}_t$. The reverse process is given as:

$$p_\theta\left(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\right)=\mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t,\boldsymbol{y}_0,t\right),\Sigma\left(\boldsymbol{x}_t,t\right)\right), \quad (2)$$

where $\mu_\theta\left(\boldsymbol{x}_t,t\right)$ is parameterized by the noise predictor $\epsilon_\theta\left(\boldsymbol{x}_t,t\right)$ and $\Sigma\left(x_t,t\right)$ is a constant dependent on α_t.

3.2. Mixture of Experts

Mixture of Experts is a modular neural network paradigm designed to enhance scalability and efficiency by dynamically activating a subset of specialized sub-models, called experts, during both training and inference. A standard MoE

Table 1. **Configurations of Diff-MoE architecture with different model sizes.** We align the Diff-MoE architecture configurations with DiT (Peebles & Xie, 2023) in S, B and L. The S+, B+ and XL configurations correspond to DiT-MoE (Fei et al., 2024). Specifically, "8E1A1S" indicates that a total of 8 sparse experts are used, with 1 expert activated for each token, and 1 expert shared by all tokens.

	#Params	#Experts	#Blocks L	Hidden Dim. D	#Head n	MLP Ratio
Diff-MoE-S	36M / 107M	8E1A1S	12	384	6	5/3
Diff-MoE-S+	77M / 211M	8E2A2S	12	384	6	4
Diff-MoE-B	137M / 402M	8E1A1S	12	768	12	5/3
Diff-MoE-B+	297M / 821M	8E2A2S	12	768	12	4
Diff-MoE-L	476M / 1.4B	8E1A1S	24	1024	16	5/3
Diff-MoE-XL	1.5B / 4.3B	8E2A2S	28	1152	16	4

system consists of N experts and a trainable gating mechanism. The gate determines the importance score for each expert based on the input $\boldsymbol{x}$, enabling sparse activation:

$$P(\boldsymbol{x}) = \text{softmax}(W_g \boldsymbol{x}), \tag{3}$$

where W_g are the gating weights. During each forward pass, only the top k experts with the highest scores are activated, and their weighted sum is computed to produce the final output. By activating only a small fraction of experts ($k \ll N$), MoE can achieve efficiently scaling without incurring the computational cost of dense architectures.

4. Methodology

In this section, we propose Diff-MoE, a novel framework that integrates Diffusion Transformers with Mixture-of-Experts, exploiting both temporarily adaptability and spatial flexiblity. We begin by enhancing the original DiT with well-established advanced architecture designs and integrate spatial MoE, replacing dense MLPs with token-routed experts to establish a high-performance baseline in Sec. 4.1. Next, we introduce expert-specific timestep conditioning in Sec. 4.2, embedding lightweight temporal signals into individual experts to dynamically adapt their behavior across denoising stages—enabling temporal adaptability with minimal parameter overhead compared to stacking separate time/space experts. Furthermore, we propose a global feature recalibration mechanism that injects contextual awareness into MoE modules in Sec. 4.3, enhancing their ability to model spatially coherent structures in image generation tasks. Finally, to address the parameter burden introduced by these two modules, we apply a low-rank decomposition technique in Sec. 4.4, ensuring the model remains efficient while incorporating the enhancements.

4.1. Basic Architecture Design

The primary objective of Diff-MoE, illustrated in Fig. 3, is to harness the powerful generative capabilities of DiT while introducing a more scalable and efficient architecture. We integrate rotary position embedding (RoPE) (Su et al., 2024) into queries and keys before self-attention—a technique proven to strengthen spatial dependency modeling in both language (Touvron et al., 2023; Liu et al., 2024a) and diffusion models (Chu et al., 2025; Tian et al., 2024; Esser et al., 2024). We also replace the standard MLP with gated linear units (GLU) (Shazeer, 2020) as feedforward network (FFN), introducing non-linearities that allow the model to learn more expressive feature representations. Inspired by previous studies (Guo et al., 2022; Li et al., 2023; Tian et al., 2024), we introduce a depthwise convolution layer prior to the MoE module, enhancing local feature extraction before channel mixing.

In addition to these architectural components, we introduce spatial Mixture-of-Experts (Dai et al., 2024), where each expert FFN shares an identical architecture. The spatial MoE enables different spatial tokens to be processed by N specialized experts, allowing the model to dynamically assign resources to different parts of the input data. This modular approach significantly improves the scalability of the model by reducing redundant computations, while still maintaining high representational capacity. The experts are sparsely activated, ensuring that computational resources are efficiently allocated according to the needs of each token. Inspired by (Dai et al., 2024; Fei et al., 2024), we incorporate additional N_s experts as shared experts, where each token will be deterministically assigned to these shared experts, avoiding the inefficient utilization of parameters caused by knowledge sharing among experts. The transformer block of Diff-MoE can be formulated as:

$$\begin{aligned} \boldsymbol{x} &= \text{MHSA}(\text{AdaLN}(\boldsymbol{x}, c)) + \boldsymbol{x}, \\ \boldsymbol{x} &= \text{MoE}(\text{AdaLN}(\text{DWConv}(\boldsymbol{x}), c)) + \boldsymbol{x}. \end{aligned} \tag{4}$$

Through the combination of these advanced architectural components and the MoE framework, Diff-MoE establishes a high-performance baseline that serves as the foundation for further innovations, discussed in later sections.

4.2. Expert-Specific Timestep Conditioning

Conventional dense diffusion models employ static architectures that uniformly process all timesteps despite the

evolving demands of the denoising process—early stages requiring coarse structural modeling and later stages fine detail refinement. This uniform approach prevents the model from dynamically adapting to the differing needs of each timestep, limiting its representational power and efficiency.

To address this issue, we introduce expert-specific timestep conditioning, building upon our spatial MoE baseline as described in Sec. 4.1. The core idea is to equip each expert with a dedicated timestep conditioning that allows it to specialize its behavior based on the current stage of the diffusion process. This enables the model to allocate processing resources more efficiently, focusing on relevant features at different stages of generation. The MoE block with expert-specific timestep conditioning can be expressed as:

$$\mathrm{MoE}(x_j) = \sum_{i=1}^{N} g_{i,j} \cdot \mathrm{AdaLN}(\mathrm{FFN}_i(x_j), c), \tag{5}$$

$$g_{i,j} = \begin{cases} P_{i,j}, & P_{i,j} \in \mathrm{Topk}(\{P_{k,t} | 1 \le k \le N\}, K), \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where x_j represents the j-th token of input x, c is the timestep condition, and $P_{i,j}$ denotes the importance score of the j-th token corresponding to the i-th expert FFN_i. This expert-specific conditioning mechanism allows each expert to specialize in processing tokens at particular timesteps, where the nature of the information to be modeled differs significantly.

In standard DiT models, time-dependent AdaLN layers create entangled routing decisions, where temporal and spatial factors compete for optimization priority. This coupling forces routers to jointly consider denoising-stage dynamics and token complexity, often resulting in suboptimal equilibria and imbalanced expert utilization. The introduction of expert-specific timestep conditioning resolves this conflict by decoupling temporal adaptation from spatial specialization. Each expert independently processes timestep embeddings through dedicated conditioning pathways, enabling routers to focus exclusively on spatial token characteristics. This architectural refinement yields two synergistic benefits: routers develop sharper discriminative capabilities for token classification while experts achieve more fine-grained timestep awareness.

4.3. Global Feature Recalibration

While MoE architectures excel at capturing local token dependencies, they often struggle with maintaining global input awareness. In typical autoregressive generative models, the absence of global context leads to suboptimal feature mixing and weak modeling of long-range dependencies. This limitation becomes particularly prominent in image generation tasks, where global coherence and local details must be balanced for high-quality outputs.

To address this, we propose a global feature recalibration mechanism specifically designed to enhance the representational capacity of expert modules in Diff-MoE. This mechanism allows each expert to incorporate global context into its processing, thereby improving the model's ability to capture complex dependencies between spatial tokens and enhancing its generative performance. The key idea behind this recalibration is the introduction of globally contextual embeddings x_g, which provide a high-level overview of the input data that is integrated into the expert's channel mixing process. This embedding enables each expert to dynamically adjust its contribution to the final output based on the broader context of the entire input sequence.

In practice, we first apply average pooling over the input tokens $\boldsymbol{x} \in \mathbb{R}^{L \times d}$ along the spatial dimension, reducing the token sequence to a global representation that captures the overall structure of the input data:

$$x_g = \mathrm{SiLU}(\mathrm{AvgPool}(\boldsymbol{x})\boldsymbol{W}_{global}), \tag{7}$$

where $\boldsymbol{W}_{global} \in \mathbb{R}^{d \times Nd}$ learnable weight matrices, SiLU represents the non-linear activation function and N is the expansion ratio of expert latent space. This step produces the globally contextual embedding $x_g \in \mathbb{R}^{1 \times Nd}$, which encodes essential global features of the input. Next, we incorporate x_g into each expert's forward pass, recalibrating the feature in high-dimensional space of GLU:

$$\mathrm{FFN}_i(x_j) = (\mathrm{SiLU}(x_j \boldsymbol{W}^i_{gate}) \cdot x_j \boldsymbol{W}^i_{up}) \boldsymbol{W}^i_{down} \cdot x_j, \tag{8}$$

where x_j is the j-th token of input x, and $\boldsymbol{W}^i_{gate} \in \mathbb{R}^{d \times Nd}$, $\boldsymbol{W}^i_{up} \in \mathbb{R}^{d \times Nd}$, $\boldsymbol{W}^i_{down} \in \mathbb{R}^{Nd \times d}$ are GLU weight matrices. The channel-wise multiplication between the expert's token processing and the global context effectively recalibrates the expert's output, enabling the model to adjust the contribution of different feature channels based on their relevance to the input distribution. By leveraging this mechanism, Diff-MoE can model complex interactions more effectively, enhancing the overall quality of generation.

4.4. Low-Rank Decomposition

To mitigate the parameter burden introduced by the proposed two modules, we apply low-rank decomposition to both components. For the expert-specific time conditioning, we factorize the temporal projection matrix $\boldsymbol{W}_t \in \mathbb{R}^{d \times Nd}$ of expert-specific AdaLN parameters (α, β, γ) into low-rank components $\boldsymbol{W}_t = \boldsymbol{A}_t \times \boldsymbol{B}_t$, where $\boldsymbol{A}_t \in \mathbb{R}^{d \times r}$ and $\boldsymbol{A}_t \in \mathbb{R}^{r \times 3d}$ with rank $r \ll d$. Similarly, the global feature recalibration weights $\boldsymbol{W}_{global} \in \mathbb{R}^{d \times Nd}$ are decomposed as $\boldsymbol{W}_{global} = \boldsymbol{A}_{global} \times \boldsymbol{B}_{global}$, retaining only the dominant singular directions.

Table 2. **Quantitative Comparison with Dense DiTs on ImageNet 256×256 dataset with $cfg = 1.0$.**

ImageNet 256×256, 400K, cfg=1.0			
Model	#Params	FID↓	IS↑
DiT-S/2	33M	66.59	20.68
SiT-S/2	33M	58.15	24.72
SiT-LLaMA-S/2	33M	53.90	26.74
Diff-MoE-S/2	36M / 107M	**44.27**	**34.27**
DiT-B/2	131M	42.84	33.66
SiT-B/2	131M	35.54	42.33
SiT-LLaMA-B/2	131	29.53	50.13
Diff-MoE-B/2	137M / 402M	**24.88**	**59.57**
DiT-L/2	458	23.27	59.63
SiT-L/2	458	19.34	70.47
SiT-LLaMA-L/2	458	14.32	86.85
Diff-MoE-L/2	476M / 1400M	**13.98**	**90.72**

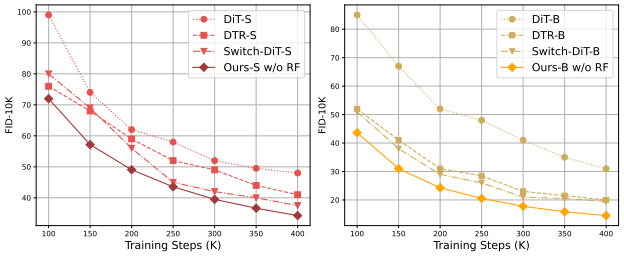


Figure 4. **Convergence comparison across different sizes.** We compare the FID-10K at different training steps for expert-based DiT models on ImageNet 256 generation with $cfg = 1.5$. For fair comparision, we report our results trained without rectified flow.

5. Experiments

5.1. Experimental Settings

Model Configurations. Diff-MoE is scaled across six configurations (S/B/L/XL and their Plus variants) to align with established Diffusion Transformer (DiT) benchmarks. As shown in Table 1, the base configurations (S/B/L) retain DiT's core hyperparameters—including 12 transformer blocks for S/B and 24 blocks for L—but replace dense MLPs with sparse MoE layers. Following previous MoE designs (Dai et al., 2024; Fei et al., 2024), we employ 8 task-specific experts by default and introduce shared experts to mitigate knowledge redundancy. To maintain parity with DiT's activated parameter count, only 1 task-specific expert and 1 shared expert are activated per token, while reducing the MLP ratio from 4 to 5/3. For example, Diff-MoE-S activates 36M parameters (107M total) under the 8E1A1S configuration (8 experts, 1 activated per token, 1 shared expert), achieving a 3× reduction in active parameters versus dense counterparts. The Plus variants expand capacity by reinstating DiT's original MLP ratio (4) and activating 2 experts per token with 2 share experts (8E2A2S). Diff-MoE-B+, for instance, scales to 821M total parameters (297M activated), preserving computational tractability while matching the scale of prior MoE-based DiT architectures (Fei et al., 2024). This hierarchical scaling ensures adaptability across hardware constraints, with the XL configuration (4.3B total parameters, 1.5B activated) targeting high-performance generation.

Table 3. **Quantitative Comparison with Time-aware Sparse DiTs on ImageNet 256×256 dataset with $cfg = 1.5$.** Specifically, "w/o RF" indicates training without rectified flow.

ImageNet 256×256, 400K, cfg=1.5			
Model	#Activated Params	FID↓	IS↑
DTR-S/2	33M	37.43	38.97
Switch-DiT-S/2	36M	33.99	42.99
Diff-MoE-S/2 w/o RF	36M	**31.18**	**50.55**
DTR-B/2	131M	16.58	87.94
Switch-DiT-B/2	144M	16.21	88.14
Diff-MoE-B/2 w/o RF	137M	**11.97**	**104.66**

Implementation Details. We utilized a pre-trained variational autoencoder (VAE) model from Stable Diffusion (Rombach et al., 2022) to encode the image and decode the latent codes. By inputting images of 256 × 256 × 3, we obtain a latent representation with dimensions of 32 × 32 × 4. We conduct experiments on class-conditional generation tasks using the ImageNet dataset (Deng et al., 2009), which contains 1,281,167 training images across 1,000 distinct classes. We train all sizes of Diff-MoE for 400k iterations using the AdamW optimizer with a learning rate of 1e-4. All models are trained with a batch size of 256. Following prior work (Park et al., 2023; Peebles & Xie, 2023), we apply exponential moving average (EMA) to the model parameters during training, with a decay factor of 0.9999, to enhance stability. Rectified flow and expert load balance loss are used by default, with further details provided in the supplementary material.

Evaluation Metrics. We evaluate image generation quality using Fréchet Inception Distance (FID) (Heusel et al., 2017), the standard metric for assessing sample quality and diversity in diffusion models. Adhering to the evaluation protocol from (Dhariwal & Nichol, 2021), FID scores are calculated over 50K generated samples with 250 DDPM sampling steps. To complement this primary assessment, we additionally report Inception Score (IS) (Salimans et al., 2016) as a secondary diagnostic of sample fidelity.

5.2. Comparison with State-of-the-Arts

Comparison with Dense DiTs. We perform a comprehensive comparison between Diff-MoE and dense DiT architectures across multiple model scales, including DiT (Fei et al.,

Figure 5. **Samples generated by Diff-MoE at 400K iterations.** We show selected class-conditional generated samples from Diff-MoE-XL/2 trained on ImageNet at 256×256 resolution.

2024), SiT (Ma et al., 2024), and SiT-LLaMA (Chu et al., 2024). DiT pioneered the integration of diffusion models with transformers, while SiT enhanced performance through improved prediction mechanisms and sampling strategies. SiT-LLaMA extends this progression by adopting a LLaMA-inspired architecture to address diverse image generation tasks. All models were trained for 400K iterations and employed classifier-free guidance (CFG) with a scale of 1.0 during the sampling phase. As shown in Table 2, Diff-MoE consistently outperforms dense counterparts at equivalent activation parameter counts. Notably, Diff-MoE-S reduces FID by 17.8% (53.90 → 44.27) and increases IS by 28.2% (26.74 → 34.27) compared to state-of-the-art SiT-LLaMA-S. Scaling to base scale models, Diff-MoE-B achieves a 15.7% IS improvement (50.13 → 59.57) and 15.7% FID reduction (29.53 → 24.88) over SiT-LLaMA-B. At large scale, Diff-MoE further lowers FID from 14.32 to 13.98 (+3.0% improvement) and elevates IS from 86.85 to 90.72 (+4.5% gain). These results demonstrate Diff-MoE's ability to enhance both image quality (FID) and diversity (IS) while maintaining parameter efficiency.

Comparison with Expert-based DiTs. We extend our evaluation beyond dense DiT models to include expert-based architectures that employ conditional computation strategies, as illustrated in Fig. 1. Current approaches specialize experts through either temporal decomposition (denoising stage selection) (Park et al., 2024; 2023) or spatial adaptation (token-wise routing) (Fei et al., 2024). DiT-MoE (Fei et al., 2024) implements spatial specialization using Top-2 routing across 10 experts (2 shared + 8 task-specific), while DTR (Park et al., 2023) applies temporal windowing to activate channel-specific experts. Switch-DiT (Park et al., 2024) explicitly assigns different denoising steps to dedicated expert modules. Unlike these dimension-specific approaches, our proposed Diff-MoE jointly optimizes expert selection across both temporal and spatial axes. To ensure fair comparison, we construct two model families with matched activation parameters: one aligned with temporal MoE baselines and another with spatial MoE counterparts. All models follow standardized training protocols (Park et al., 2024) (400K iterations) and sampling parameters (CFG scale 1.5). Table 3 demonstrates Diff-MoE-S's superiority over tempo-

Table 4. **Quantitative Comparison with Space-adaptive Sparse DiTs on ImageNet 256×256 dataset with $cfg = 1.5$.**

ImageNet 256×256, 400K, cfg=1.5			
Model	#Activated Params	FID↓	IS↑
DiT-MoE-S/2	71M	17.92	80.50
Diff-MoE-S/2+	77M	**9.27**	**126.22**
DiT-MoE-B/2	286M	7.19	147.67
Diff-MoE-B/2+	297M	**4.00**	**195.76**
DiT-MoE-XL/2	1.5B	3.42	214.73
Diff-MoE-XL/2	1.5B	**2.69**	**262.22**

Table 5. **Ablation Study of Basic Architecture Design on ImageNet 256×256 dataset with $cfg = 1.5$.**

ImageNet 256×256, 400K, cfg=1.5			
Model	#Params	FID↓	IS↑
DiT-S/2	32.96M	43.02	35.80
+ RoPE 2D	32.96M	39.80	39.10
+ GLU	32.94M	38.20	41.43
+ DepthWise Conv	33.06M	35.85	42.85
+ Spatial MoE	38.37M	32.74	47.19

ral MoE baselines, achieving a 12.3% FID reduction versus DTR-S and 8.3% improvement over Switch-DiT-S. This performance gap amplifies with model scale, confirming our method's enhanced scalability. Complementary comparisons with spatial MoE architectures in Table 4 reveal consistent superiority over DiT-MoE variants. Most notably, Diff-MoE-XL/2 achieves a 27% FID reduction (3.42 → 2.69) compared to DiT-MoE-XL/2, while maintaining similar number of parameters. The accelerated convergence rate of Diff-MoE, evidenced by training dynamics in Fig. 4, underscores the efficiency of our spatiotemporal paradigm. These consistent improvements across architectural scales and optimization timelines highlight the effectiveness of joint spatial-temporal expert coordination.

5.3. Ablation Study

Basic Architecture Design. Building upon the foundation of DiT, we have made the following improvements to the network architecture: 1. Rotary Position Embedding (RoPE): Integrated into self-attention queries/keys to strengthen spatial awareness; 2. Gated Linear Unit (GLU): Replaces standard FFN MLP layers to enhance nonlinear feature learning; 3. Depthwise Convolution layers before the expert model: Boosting local feature extraction. We quantitatively validated the effectiveness of these modifications on the conditional generation task using the ImageNet dataset. Quantitative evaluations on ImageNet conditional generation (Table 5) demonstrate progressive improvements: RoPE integration reduces FID from 43.02 to 39.80 (+7.5% improvement), GLU substitution further lowers FID to 38.20 (+4.1% gain), while depthwise convolutions achieve FID 35.85 (+6.5% improvement) with minimal parameter overhead (+0.8%). The complementary integration of spatial MoE yields a final FID of 32.74 (+9.3% reduction).

Table 6. **Ablation Study of MoE Design on ImageNet 256×256 dataset with $cfg = 1.5$.**

ImageNet 256×256, 400K, cfg=1.5			
Model	#Acti. Params	FID↓	IS↑
MoE Baseline	38.37M	32.74	47.19
+ Expert-Spec. Time Cond.	49.02M	28.71	54.57
+ Global Feat. Recalibration	53.74M	26.33	59.54
+ Low Rank	41.84M	27.59	55.62

MoE Design. We systematically evaluate the core components of Diff-MoE's spatiotemporal expert architecture through controlled ablations. As evidenced by Tab. 6, integrating expert-specific timestep conditioning into the spatial MoE framework substantially reduces FID from 32.74 to 28.71, confirming the necessity of modeling denoising-stage dynamics in expert design. Further performance gains emerge from our global feature calibration module, which lowers FID to 26.33 by injecting global context into locally processed tokens—demonstrating that global guidance enhances local feature coherence despite expert specialization. To ensure parameter efficiency comparable to baseline DiT models, we implement low-rank decomposition techniques (r=64) that reduce total parameters by 28%. This optimization incurs a moderate performance trade-off (FID 27.59), yet still maintains superior results over standard spatial MoE implementations. The residual performance gap highlights our architecture's effectiveness in balancing capacity and efficiency.

6. Conclusion

In this paper, we presented Diff-MoE, a novel framework that integrates Diffusion Transformers with Mixture-of-Experts (MoE) to address the scalability and representational limitations of existing diffusion models. Our architecture introduces timestep-aware experts that dynamically adapt to denoising stages—specializing in coarse semantic shaping during early diffusion steps and fine-grained detail refinement in later stages—paired with space-adaptive experts that perform localized feature processing through token-level routing. A global feature recalibration mechanism further enhances representational power of expert modules by adapting feature contributions based on their relevance to the input. Our framework outperforms both dense DiT and existing MoE-based methods, offering a robust solution for high-performance generative modeling.

Limitation. While our experiments empirically validate

Diff-MoE's efficacy at 400K training iterations, we acknowledge that computational resource constraints preclude evaluation of large-scale variants (e.g., Diff-MoE-XL and DiT alignment protocols at 7M steps). This limitation leaves open the full exploration of our method's potential at extreme scales, which is left as our future work.

Impact Statement

This paper presents work whose goal is to advance the field of image generation. Similar to other image generation methods, our approach must be used cautiously to prevent potential misuse.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants U22A2096 and 62036007, in part by Scientific and Technological Innovation Teams in Shaanxi Province under grant 2025RS-CXTD-011, in part by the Shaanxi Province Core Technology Research and Development Project under grant 2024QY2-GJHX-11, in part by the Fundamental Research Funds for the Central Universities under GrantQTZX23042, in part by the Innovation Fund of Xidian University.

References

Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S., et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., and Li, Z. Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024.

Chu, X., Su, J., Zhang, B., and Shen, C. Visionllama: A unified llama backbone for vision tasks. In *European Conference on Computer Vision*, 2024.

Chu, X., Su, J., Zhang, B., and Shen, C. Visionllama: A unified llama backbone for vision tasks. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2025.

Dai, D., Deng, C., Zhao, C., Xu, R., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Fei, Z., Fan, M., Yu, C., Li, D., and Huang, J. Scaling diffusion transformers to 16 billion parameters. *arXiv preprint arXiv:2407.11633*, 2024.

Feng, Z., Zhang, Z., Yu, X., Fang, Y., Li, L., Chen, X., Lu, Y., Liu, J., Yin, W., Feng, S., et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10135–10145, 2023.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., and Xu, C. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12175–12185, 2022.

Hatamizadeh, A., Song, J., Liu, G., Kautz, J., and Vahdat, A. Diffit: Diffusion vision transformers for image generation. In *European Conference on Computer Vision*, pp. 37–55. Springer, 2024.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Huang, Q., An, Z., Zhuang, N., Tao, M., Zhang, C., Jin, Y., Xu, K., Chen, L., Huang, S., and Feng, Y. Harder tasks need more experts: Dynamic routing in moe models. *arXiv preprint arXiv:2403.07652*, 2024.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35: 26565–26577, 2022.

Labs, B. F. Flux. https://github.com/black-forest-labs/flux, 2023.

Lee, Y., Kim, J., Go, H., Jeong, M., Oh, S., and Choi, S. Multi-architecture multi-expert diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13427–13436, 2024.

Li, Y., Zhang, K., Cao, J., Timofte, R., and Van Gool, L. Localvit: bringing locality to vision transformers (2021). *arXiv preprint arXiv:2104.05707*, 2023.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Liu, Q., Zeng, Z., He, J., Yu, Q., Shen, X., and Chen, L.-C. Alleviating distortion in image generation via multi-resolution diffusion models and time-dependent layer normalization. *Advances in Neural Information Processing Systems*, 37:133879–133907, 2024b.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Ma, N., Goldstein, M., Albergo, M. S., Boffi, N. M., Vanden-Eijnden, E., and Xie, S. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*, 2024.

Mei, K. and Patel, V. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9117–9125, 2023.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Park, B., Woo, S., Go, H., Kim, J.-Y., and Kim, C. Denoising task routing for diffusion models. *arXiv preprint arXiv:2310.07138*, 2023.

Park, B., Go, H., Kim, J.-Y., Woo, S., Ham, S., and Kim, C. Switch diffusion transformer: Synergizing denoising tasks with sparse mixture-of-experts. *arXiv preprint arXiv:2403.09176*, 2024.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., and Chen, Q. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15932–15942, 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. 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*, pp. 234–241. Springer, 2015.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Shazeer, N. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. The sparsely-gated mixture-of-experts layer. *Outrageously large neural networks*, 2017.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Sun, H., Lei, T., Zhang, B., Li, Y., Huang, H., Pang, R., Dai, B., and Du, N. Ec-dit: Scaling diffusion transformers with adaptive expert-choice routing. *arXiv preprint arXiv:2410.02098*, 2024.

Tian, Y., Tu, Z., Chen, H., Hu, J., Xu, C., and Wang, Y. U-dits: Downsample tokens in u-shaped diffusion transformers. *arXiv preprint arXiv:2405.02730*, 2024.

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. *arXiv preprint arXiv:2302.13971*, 2023.

Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., and Zhu, J. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024.

Xue, Z., Song, G., Guo, Q., Liu, B., Zong, Z., Liu, Y., and Luo, P. Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems*, 36, 2024.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024b.

Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A. M., Le, Q. V., Laudon, J., et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

Zoph, B., Bello, I., Kumar, S., Du, N., Huang, Y., Dean, J., Shazeer, N., and Fedus, W. Designing effective sparse expert models. *arXiv preprint arXiv:2202.08906*, 2(3): 17, 2022.

Zuo, S., Liu, X., Jiao, J., Kim, Y. J., Hassan, H., Zhang, R., Zhao, T., and Gao, J. Taming sparsely activated transformer with stochastic experts. *arXiv preprint arXiv:2110.04260*, 2021.

A. Diff-MoE Performance Summary

Table 7. **Performance of Diff-MoE architecture with different model sizes.**

ImageNet 256×256, 400K, cfg=1.5						
	#Params	FID↓	sFID↓	IS↑	Precision↑	Recall↑
Diff-MoE-S/2	36M / 107M	17.94	83.67	6.56	0.6472	0.5401
Diff-MoE-S/2+	77M / 211M	9.27	126.22	5.72	0.7310	0.5285
Diff-MoE-B/2	137M / 402M	6.45	155.81	5.26	0.7629	0.5373
Diff-MoE-B/2+	297M / 821M	4.00	195.76	4.88	0.7990	0.5447
Diff-MoE-L/2	476M / 1.4B	3.12	223.51	4.79	0.8162	0.5473
Diff-MoE-XL/2	1.5B / 4.3B	2.69	262.22	4.54	0.8364	0.5579

B. Training

B.1. Rectified Flow

Rectified Flow (RF) (Liu et al., 2022) reformulates generation as a deterministic ordinary differential equation (ODE). It learns a straight trajectory in probability space, directly mapping noise $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I})$ to data $\boldsymbol{x}_0$ via a velocity field $\boldsymbol{v}_\theta$:

$$\frac{d\boldsymbol{x}_t}{dt} = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t), \quad t \in [0, 1], \tag{9}$$

where $\boldsymbol{x}_t = (1-t)\boldsymbol{z} + t\boldsymbol{x}_0$ enforces a linear interpolation. The training loss minimizes trajectory curvature:

$$\mathcal{L}_{\text{rf}} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{z}} \left| \boldsymbol{v}_\theta \left((1-t)\boldsymbol{z} + t\boldsymbol{x}_0, t \right) - (\boldsymbol{x}_0 - \boldsymbol{z}) \right|^2. \tag{10}$$

We present a convergence analysis comparing the training dynamics of DDPM-based and Rectified Flow-based implementations of our Diff-MoE-S architecture, both trained for 400,000 iterations on ImageNet 256×256 generation tasks under classifier-free guidance (CFG scale=1.5). As shown in Fig.6, the rectified flow-based model converges faster than the DDPM-based model. For context, Rectified Flow is used in SiT, SiT-Llama, and DiT-MoE, while DDPM is applied to other methods. To ensure a fair comparison, we report results without rectified flow in Tab.3 and Fig. 4.

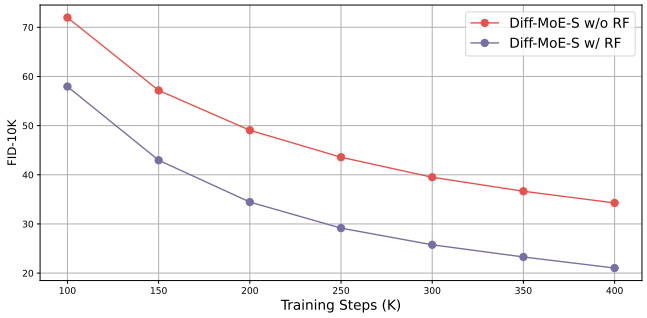


Figure 6. **Convergence comparision of DDPM and Rectified Flow-based Diff-MoE-S.** We compare the DDPM-based Diff-MoE-S and Rectified Flow-based Diff-MoE-S trained for 400K iterations on ImageNet 256×256 generation with $cfg = 1.5$.

B.2. Load Balance Loss

Building on insights from prior work (Zuo et al., 2021) demonstrating that uniform expert utilization in MoE layers correlates with enhanced model performance, we incorporate a differentiable load-balancing loss (Fedus et al., 2022) to enforce equitable workload distribution. For a MoE layer with N experts processing L input tokens, the load-balance loss can be calculated as:

$$\mathcal{L}_{load} = \alpha \cdot N \cdot \sum_{i=1}^{N} f_i \cdot P_i, \tag{11}$$

where α controls the loss magnitude, and the expert-specific terms are computed as:

$$f_i = \frac{1}{L} \sum_{j=1}^{L} \mathbb{I}\left\{\arg\max p(x) = i\right\}, \tag{12}$$

$$P_i = \frac{1}{L} \sum_{j=1}^{L} P_{i,j}(x). \tag{13}$$

Here $P(x)$ denotes the routing probability distribution from Eq. 3, with f_i representing the empirical fraction of tokens assigned to expert i, and P_i is the corresponding mean routing probability. This formulation penalizes discrepancies between actual token assignments and the router's probability mass allocation.

B.3. Final Loss

Our model is a diffusion model that utilizes either diffusion loss (Ho et al., 2020) or velocity field loss in Eq. 10 as the generative training objective. The final loss function combines the generative loss with the load-balance loss, as follows:

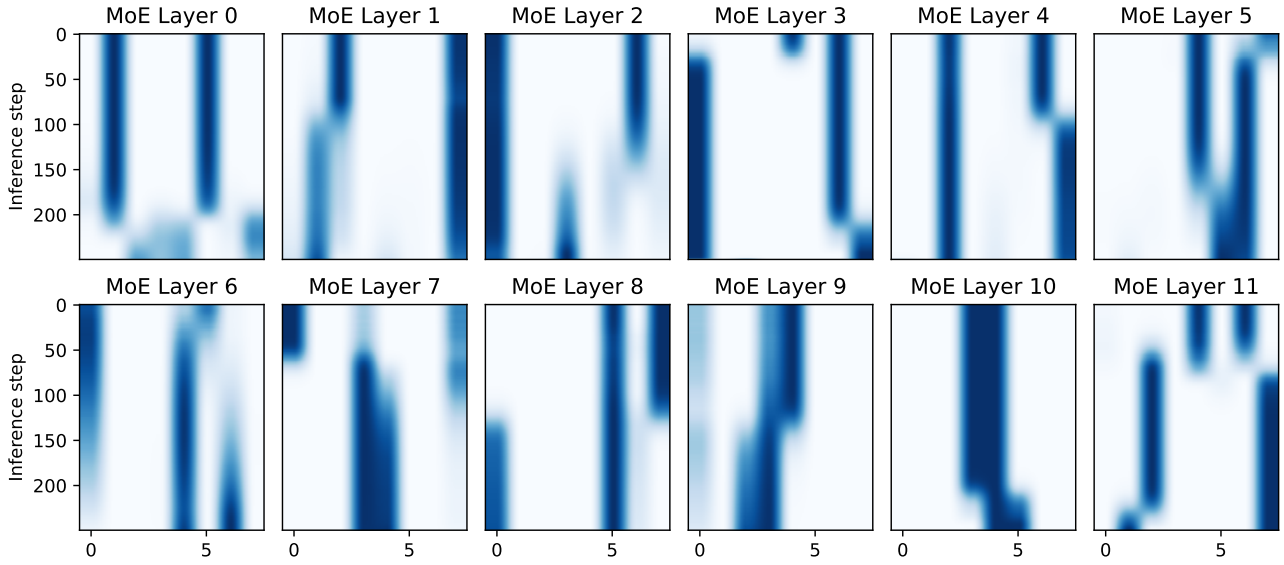
$$\mathcal{L}_{final} = \begin{cases} \mathcal{L}_{ddpm} + \mathcal{L}_{load}, & \text{DDPM}, \\ \mathcal{L}_{rf} + \mathcal{L}_{load}, & \text{Rectified Flow}. \end{cases} \tag{14}$$

C. Expert Selection Analysis

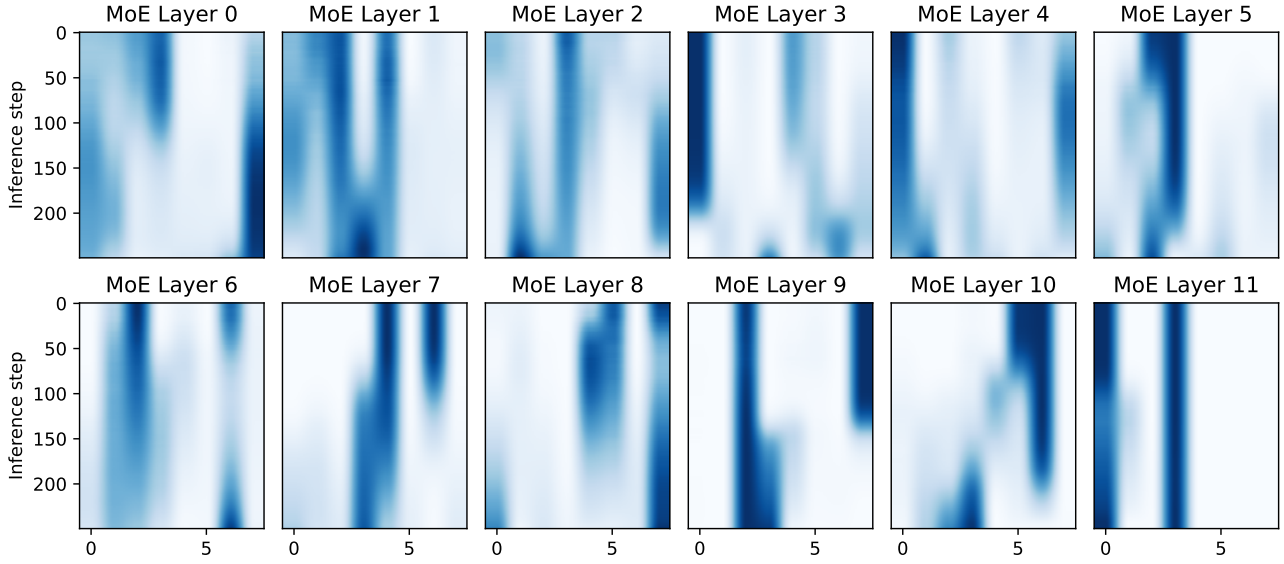
We analyze the routing dynamics of Diff-MoE-S/2 through comparative visualizations of its 12 MoE layers (Fig. 7), contrasting configurations with and without Expert-Specific Time Conditioning. While this mechanism directly modulates expert parameters rather than router weights, its indirect influence on routing behavior emerges as a critical finding.

In MoE baseline (first variant of Tab. 6), time-dependent AdaLN layers in the main DiT path create entangled routing decisions where temporal and spatial factors compete for optimization priority. This coupling forces routers to jointly account for denoising-stage dynamics and token complexity, often converging to suboptimal equilibria characterized by imbalanced expert utilization.

The introduction of Expert-Specific Time Conditioning resolves this conflict by decoupling temporal adaptation from spatial specialization. Each expert independently processes timestep embeddings through dedicated conditioning pathways, enabling routers to focus exclusively on spatial token characteristics. This architectural refinement yields two synergistic benefits: routers develop sharper discriminative capabilities for token classification while experts achieve more balanced workload distributions. Visual evidence in Fig. 7 confirms these improvements, showing reduced expert activation variance and clearer spatial feature boundaries compared to the unconditioned baseline.



(a) w/o Expert-Specific Time Conditioning



(b) w/ Expert-Specific Time Conditioning

Figure 7. **Frequency for selected experts per timestep.** We visualize the 12 MoE layers of the Diff-MoE-S/2 variant, both with and without Expert-Specific Time Conditioning. The x-axis represents the 8 experts in each layer, while the y-axis corresponds to the 250 DDPM steps used to sample the synthesized image. For each pair (expert e, inference step i), we show the average routing frequency across all timesteps i assigned to that specific expert e.

D. More Visualization Results

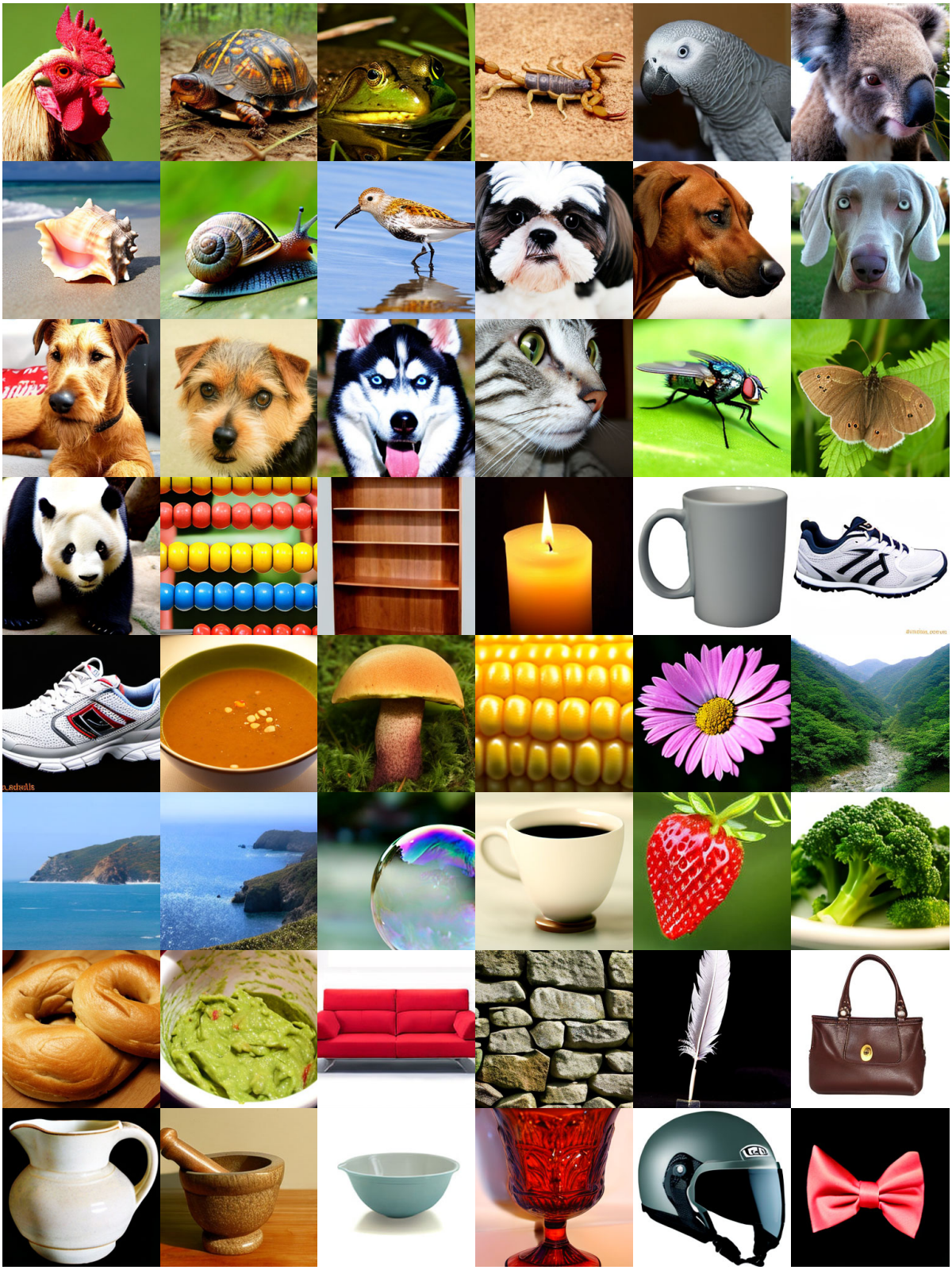


Figure 8. More Samples generated by Diff-MoE-XL/2 trained on IamgeNet 256 at 400K iterations.