Mixture of States: Routing Token-Level Dynamics for Multimodal Generation

Supplementary Material

1. Related Work

Text-to-Image Network Architecture Diffusion models [33, 40, 56, 61] have become a dominant paradigm for text-to-image generation [5, 12, 15, 19, 66, 68, 69] owing to their scalability and stable training. In multimodal diffusion models, prompt embeddings derived from a frozen text encoder are incorporated through cross-attention [12, 62, 68], self-attention [9, 21], or layer-wise attention (e.g., MoT) [47]. A fundamental challenge in this design is the "static vs. dynamic" mismatch. The diffusion process is inherently dynamic, operating over numerous timesteps with varying noise levels and visual features [42, 52]. However, the text encoder provides only a single, static representation of the prompt. While self- and layer-wise attention allow this conditional information to evolve within the visual backbone's blocks, the initial conditional signal provided to the model remains fixed. To address this limitation, our MoS framework introduces a learnable router. The router jointly considers the prompt, denoising step, and noised image to dynamically select and aggregate conditional embeddings, enabling true input- and time-dependent conditioning.

Unified Model Recent research [24, 75, 81, 86] has increasingly sought to unify diverse tasks within a single framework. While following this unified design philosophy, it diverges in its training methodology. Instead of the common approach of jointly training all tasks in a single, complex stage [13, 18, 25, 48, 87, 95], we adopt a multi-stage training strategy. This approach provides significant flexibility and efficiency. Specifically, by freezing our text branch, we can focus computational resources purely on optimizing the multimodal generation components. This staged training also circumvents common challenges of joint training, such as throughput bottlenecks from mixed data batches and the difficulty of balancing diverse, and often competing, learning objectives across modalities. This strategy is consistent with other recent, successful large-scale models like BLIP-3o [11], MetaQuery [60], Qwen-Image [79] and LMFusion [72], which also employ multi-stage training.

Dynamic Neural Networks MoS is also related to the principles of dynamic neural networks [29, 31, 37–39, 70], in which the computational graph or parameter usage is conditioned on the input. A prominent example is the Mixture-of-Experts (MoE) [14, 20, 22, 29, 41, 44, 54, 71], where tokens are adaptively processed by different "expert" sub-networks within each transformer block. Due to its sparsity, MoE efficiently scales model parameters while keeping computation tractable. Recent advances have explored other forms of dynamic computation, such as Mixture-of-Depths (MoD) [67], which dynamically allocates compute across tokens and layers, and Mixture-of-Recursions (MoR) [2], which reuses layers recursively with variable-depth routing. These methods establish a powerful principle: computation should be sparse, adaptive, and conditional on the input. However, this principle has largely been applied to *intra-model adaptivity*, i.e., routing tokens within a single large model. In contrast, MoS extends this principle to *inter-model collaboration*.

Mixture of Transformers (MoT) Our model, MoS, is conceptually related to the Mixture of Transformers (MoT) architecture [47]. In MoT, each modality is processed by an independent transformer, while a shared attention module in each block enables tokens to attend across modalities. This design has been widely adopted in multimodal research: LMFusion [72] and PGV3 [51] use it to couple a frozen LLMs as the text encoder with a trainable diffusion transformer for strong text-to-image generation performance. More recently, Bagel [18] and Mogao [48] enable joint training of both transformers, unifying image understanding and generation. Despite these advances, a critical limitation persists: prior MoT variants require identical hidden dimensions and a strict one-to-one block correspondence across modalities to share global attention. This rigid, symmetric constraint is highly inflexible, as various modalities may follow distinct scaling laws and design principles. MoS is designed to solve this specific problem. We replace the rigid global attention mechanism with a learnable, sparse router, which removes the identical-size constraint and enables adaptive, effective interactions between asymmetric transformers.

2. Implementation Details

To support efficient and stable training, we introduce a set of optimizations spanning both system-level infrastructure and model design:

- *QK-Norm:* To enhance training stability, we apply QK-Norm [32] in each transformer block. Specifically, before the attention operation, we normalize the query and key vectors using RMS-Norm [91].
- *Modality-specific Norm:* We apply separate normalization layers for different modalities, which improves performance while maintaining training stability.

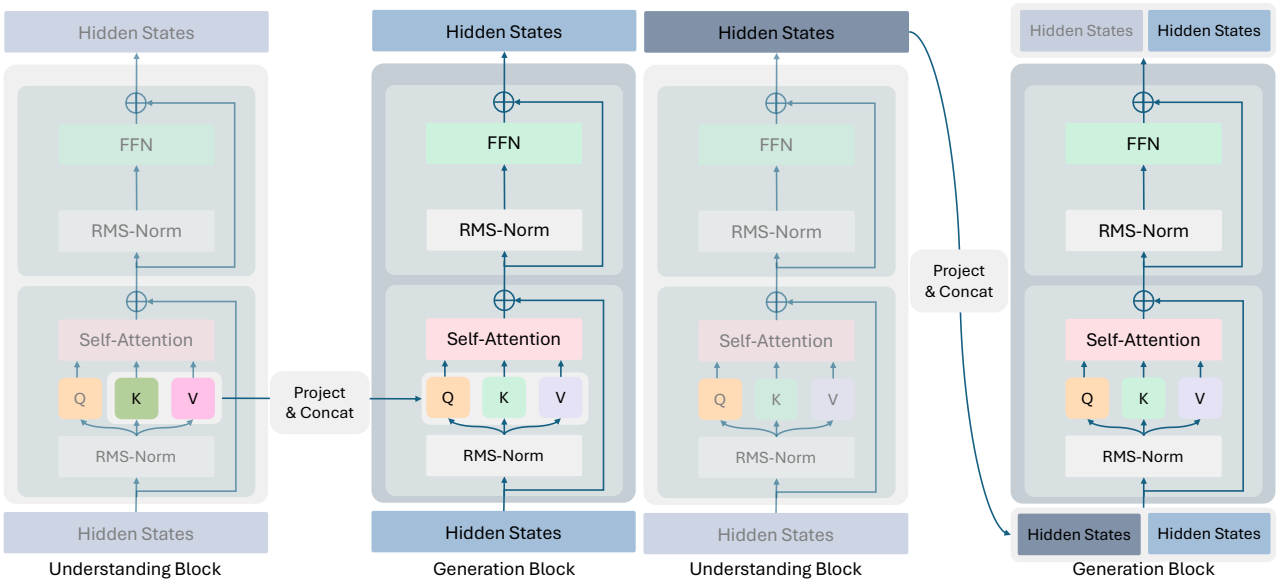


Figure 1. **Illustration of the router's operation space.** (a) In the global-attention scheme, keys and values from the understanding branch are projected to the generation branch for cross-modal interaction. (b) In the global hidden-state scheme, hidden representations from both branches are directly concatenated at the input of each transformer block.

- *Token Registers:* Following Darcet et al. [16], we introduce four auxiliary learnable tokens into the input sequence to enhance training stability, without assigning them explicit training objectives.
- *Diffusion Step Sampling:* To accelerate convergence during low-resolution (512×512) pretraining, we adopt logit-normal sampling, which is later replaced by mode sampling (scale = 0.8, shift = 3.0) to adapt to high-resolution (1024×1024 and 2048 × 2048) training.
- *Dropping timestep embeddings.* Motivated by recent findings [73, 74] and the empirical analysis, we confirm that timestep embeddings provide negligible benefit to the diffusion model while introducing an overhead of ∼20% parameters. For efficiency, we remove the timestep conditioning from the generation tower.
- *FSDP:* We adopt Fully Sharded Data Parallel (FSDP) as our primary distributed training framework and enable activation checkpointing in the high-resolution stages.
- *Low-Precision Training*: We employ a module-specific mixed-precision strategy. The VAE compressor is maintained in float32 to ensure numerical stability, while the understanding tower is set entirely in bfloat16. For the generation tower and router, we use bfloat16 for all-gather operations and float32 for gradient reduce-scatter.
- *Triton Kernel Optimization:* To further improve training throughput, we employ custom Triton kernels, including an RMSNorm kernel and a fused FFN kernel. Our implementation builds on the liger-kernel [34].
- *Bucket-wise Dynamic Resolution Training:* To support dynamic-resolution training, we adopt a resolution-driven bucket dataloader. Data samples are assigned online to buckets based on their resolution and aspect ratio. Once a bucket is filled, it is dispatched to the model for training.
- *Data Reweighting:* To achieve balanced performance across different dimensions, we adjust the mixing ratios of the training datasets. The optimal ratios are determined through grid search.

Regarding the hyperparameters, we use AdamW with a learning rate of 1×10^{-4}, weight decay of 0.01, and betas set to $(0.9, 0.95)$. The first 4k steps serve as a warm-up phase, where the learning rate is linearly increased to the target value. A cosine schedule is then applied to gradually decay the learning rate to 1.5×10^{-5}. The global batch size is dynamically set to 2048 or 1024, depending on available training resources. For each pre-training stage, we run 400k–1200k steps based on visual inspection of convergence, while HQ fine-tuning is performed for 50k steps. We use a top-k router ($k = 2$) with ϵ-greedy exploration ($\epsilon = 0.05$). The training platform comprises both A100 and H100 GPUs; to standardize reporting, we compute the total training cost by counting 2 A100 days as equivalent to 1 H100 day.

3. Additional Discussions on Router Designs

3.1. Router Operation Space

This ablation study aims to determine the MoS router's operation space, i.e., which features routed across transformers yield the optimal benefit. We begin with a solid baseline—the MoT architecture [47]. To enable representation transfer across transformers, MoT introduces a global attention mechanism where keys, values, and queries are shared between towers. In contrast to MoT, a competing approach fuses hidden states directly before sequence modeling. As shown in Fig. 1 (a)-(b), these two design philosophies yield four candidates:

- *Global Attention (Head-Projection)*[1]: Apply a projection layer on each head dimension for the key and value vectors from the understanding tower, then concatenate with the generation tower's key and value representations, respectively.
- *Global Attention (State-Projection)*: Apply the projection layer on the hidden dimension, split into multi-head vectors, and fuse with the generation tower's features.
- *Global Hidden States (Independent-Projection)*: Concatenate hidden states in the generation tower, then apply separate key-value projection layers before attention.
- *Global Hidden States (Shared Projection)*: Concatenate hidden states, then apply a shared key–value projection layer.

As shown in Fig. 2 (a)-(b), the empirical results clearly indicate that using global hidden states with a shared projection layer yields the optimal configuration. Note that the MoS router is excluded from this ablation.

3.2. Router Architecture Design

In the MoS router, token embeddings from different modalities are normalized using separate RMSNorm layers to align their representation scales. Our analysis shows that this design is a key factor in improving performance. To verify its effectiveness, we compare two configurations: (a) a shared RMSNorm applied to all modalities, and (b) separate RMSNorms for each modality. Fig. 2(c–d) shows that configuration (b) consistently outperforms (a) across all evaluation metrics.

3.3. ϵ-greedy Strategy and Sparsity Design

To evaluate the impact of applying ϵ-greedy to the router's output, we conduct an ablation study. Here, $\epsilon = 0.05$, meaning that with a 5% probability the router randomly selects a layer rather than following the predicted logits. This choice is informed by an empirically grid search. As shown in Fig. 3, the results indicate that incorporating ϵ-greedy notably accelerates convergence across training steps. Next, we study how many layers (k) should be consolidated to form the final guidance feature. As shown in Fig. 3, $k = 2$ consistently outperforms other candidates. This is reasonable: when $k = 1$, the model may become trapped in a local view, as the router tends to overfit to a single layer; conversely, larger k values dilute representations, over-flatten hidden states, and ultimately degrade performance.

3.4. Scalability of MoS

Here, we validate the scalability of our model. Prior studies [7, 21, 64, 79] have demonstrated the effectiveness of scaling the generation tower (diffusion model). Since our approach does not alter the fundamental formulation of the diffusion process, it should likewise benefit from enlarging the generation tower. In this section, however, we focus on a complementary direction—scaling the understanding tower. Unlike MoT, MoS provides a more flexible framework that enables independent scaling of the understanding tower, thereby allowing the use of larger understanding models. As shown in Fig. 4 (a)-(b), we find that enlarging the text encoder yields consistent and stable improvements as its size increases. Moreover, since the understanding tower does not need to be updated across all training stages and its embeddings can be provided in a Producer–Consume manner, scaling the understanding tower emerges as a cost-efficient solution based on our empirical analysis.

3.5. MoS vs. Fixed Router Solutions

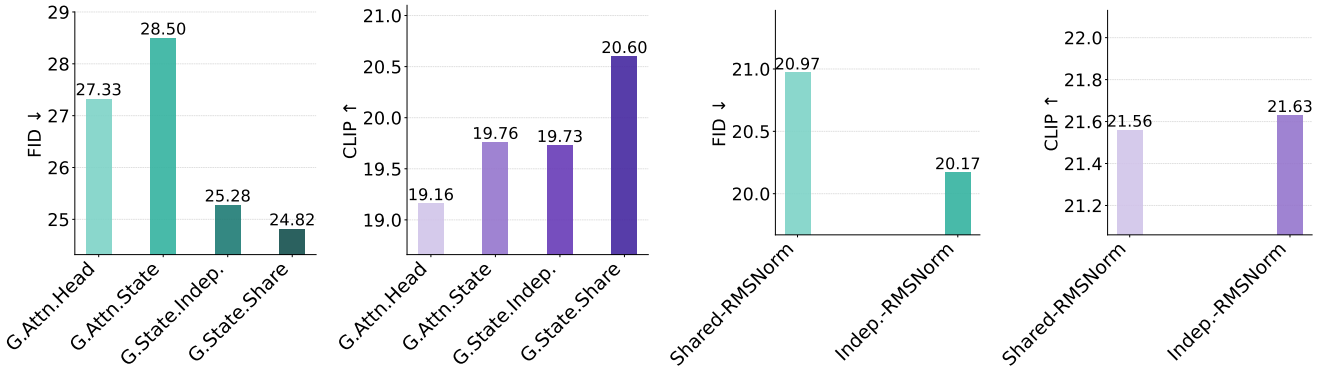
To verify the effectiveness of MoS, we compare it with several fixed-routing baselines. These baselines use either the final-layer embedding [A], a middle-layer embedding [B], or a combination of middle-layer embeddings [C] with a fixed 1:1 routing strategy. All methods are trained using identical data, model parameters, and training steps to ensure a fair comparison.

Tab. 1 shows that utilizing multi-layer hidden representations improves performance over using a single-layer embedding. In particular, the combination of middle-layer embeddings [C] consistently outperforms both the final-layer and single middle-layer variants. Nevertheless, MoS further achieves substantial improvements, indicating that dynamic routing across multiple hidden states is more effective than fixed routing strategies. We further present the performance curves comparing MoS and [C] in Fig. 4.

3.6. MoS for Image Editing

Our design feeds the reference image into both the understanding and generation towers. We hypothesize that this allows the model to leverage semantic information from the understanding tower and low-level visual features from the generation tower, thereby enabling more precise and con-

[1]Here, we avoid using the term layer-wise attention to describe this operation, to maintain symmetry with the case of global hidden states. Nonetheless, the mechanism is essentially equivalent to the layer-wise attention discussed in previous sections.



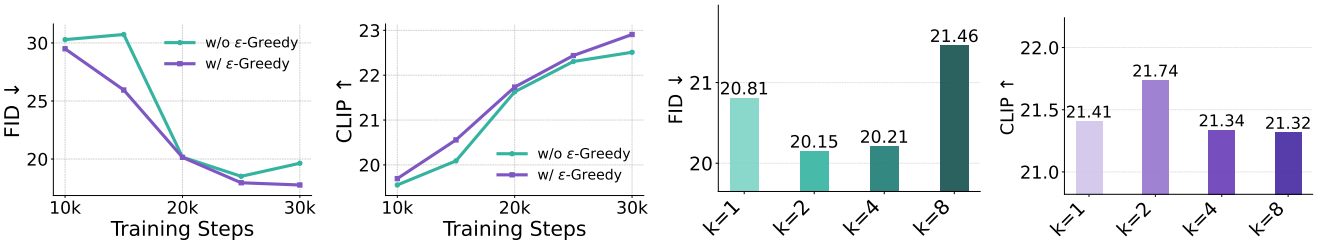
(a) FID comparison across feature types used for interaction.

(b) CLIP comparison across feature types used for interaction.

(c) FID performance with and without separate normalization in router.

(d) CLIP performance with and without separate normalization in router.

Figure 2. **Ablation study results on FID and CLIP across the router's operation space and architectural design.** (a)–(b) indicate that using hidden states as the router's operation space outperforms using key/value features, while (c)–(d) show that applying modality-specific normalization to the router's inputs further improves performance.



(a) FID performance by using ϵ-greedy.

(b) CLIP performance by using ϵ-greedy.

(c) FID performance across different top-k settings.

(d) CLIP performance across different top-k settings.

Figure 3. **Ablation study on the ϵ-greedy exploration strategy and sparsity settings.** Results show that incorporating ϵ-greedy accelerates convergence, and that $k = 2$ yields the best performance.

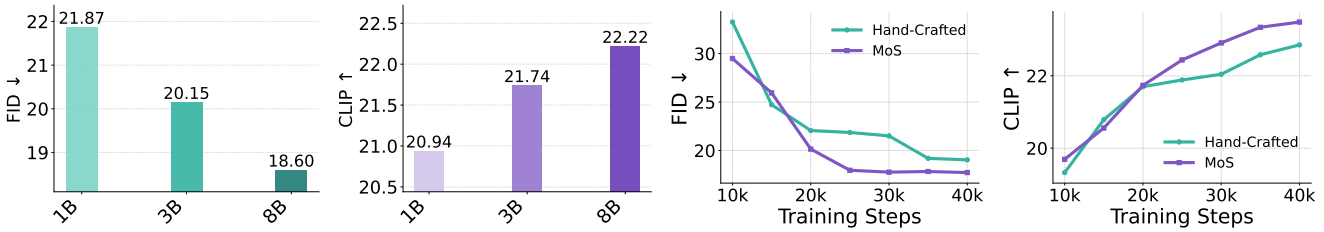
Table 1. Comparison between MoS and fixed routing baselines on the MJHQ benchmark. [A] uses the final-layer embedding for routing, [B] uses a single middle-layer embedding, and [C] uses a combination of middle-layer embeddings with a fixed 1:1 routing strategy. Results demonstrate that leveraging multi-layer hidden representations improves performance, while MoS further provides a more effective routing mechanism than fixed routing.

Method	FID-20k ↓	CLIP-20k ↑
A. Final-layer embedding	25.70	20.26
B. Middle-layer embedding	22.72	21.39
C. Multi-layer embedding (fixed routing)	22.06	21.70
MoS	**20.15**	**21.74**

sistent editing. To validate the effectiveness of our design, we compare three input configurations for image editing: (i) *w/o generation-tower context*, where the reference image input to the generation tower is removed; (ii) *w/o understanding-tower context*, where the reference image input to the understanding tower is removed; and (iii) *w/ full context*, where both towers receive the source images. As shown in Fig. 5, we conduct experiments on GEdit-Bench [55], which evaluates model editing performance across three dimensions: semantic consistency (G-SC), perceptual quality (G-PQ), and overall score (G-O). All metrics are obtained from GPT-4o–based automatic evaluations [36], where higher values indicate better performance. The results demonstrate that incorporating reference images from both towers achieves the highest average score, consistent with our hypothesis.

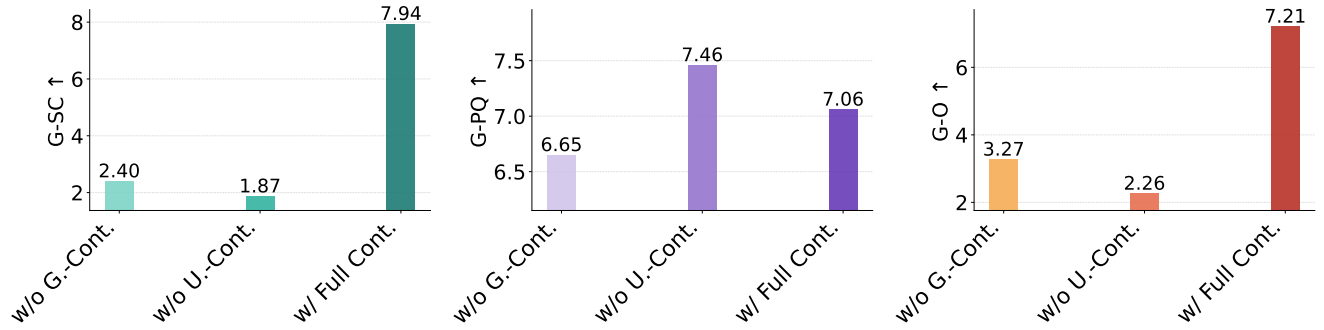
3.7. Ablation Study on Inference Strategy

MoS incorporates denoising steps into its input, which may influence the underlying diffusion dynamics. We thereby empirically validate whether common inference enhancements can benefit MoS-based models. As shown in Fig. 6, we evaluate an intermediate checkpoint of MoS-S (trained for 400k steps at 512×512 resolution) on GenEval [26]. The results show that increasing inference steps consistently improves generation quality, and CFG guidance can be applied in the typical range (5.0–7.5), similar to the other models [62]. We further observe that adopting a linear–quadratic scheduler [63] and rescaling strategy [50]



(a) FID performance by using different Und. Tower. (b) CLIP performance by using different Und. Tower. (c) FID performance across different interaction setting. (d) CLIP performance across different interaction setting.

Figure 4. **Ablation study results on FID and CLIP for understanding tower size and interaction types.** Our ablations show that MoS interaction consistently outperforms hand-crafted design, while also benefiting from scaling up the understanding tower.



(a) G-SQ (Semantic Consistency) Performance with varying context. (b) G-PQ (Perceptual Quality) Performance with varying context. (c) G-O (Overall Score) Performance with varying context.

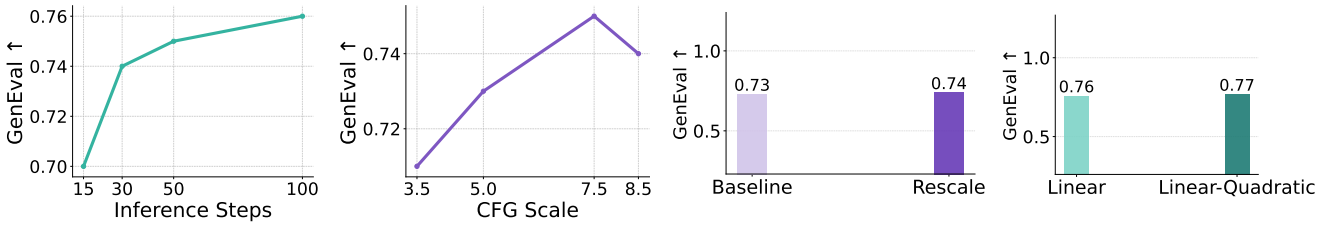
Figure 5. **Ablation study results on GEdit-Bench [55].** The best performance is obtained when the source images are provided to both the generation and editing towers.

yields slight additional gains.

3.8. MoS Router Visualization Analysis

To analyze the router's behavior, we visualize its output patterns in Fig. 7 using the caption "A dog holding a sign that says 'MoS in 2025'" with MoS-S: i) The first row shows the denoising trajectory, where the model progressively refines the image from pure noise to the target output, guided by the input caption. ii) The second row visualizes the average contribution of each understanding layer. To obtain this, we compute the router's logit matrix—modeling the affinity between blocks in the understanding and generation towers—and average the weights across all generation blocks and tokens. iii) The third row presents the router's output at a fixed denoising step (t=1) for individual tokens, revealing that different tokens induce distinct routing patterns. The results indicate that:

- The router's predictions vary across denoising steps. In the early stages, features from layers of different depths are sparsely selected as the most influential. As the denoising process progresses, the weights of the middle layers gradually increase, leading to smoother importance distributions and reduced variation across steps. This trend is intuitive: at later stages, most semantic information has been established, and the model no longer requires highly specific features from individual layers. This observation is consistent with the findings reported in [52].
- The router's predictions also vary across tokens. As shown in Fig. 7, each token exhibits a distinct connection pattern, reflecting the router's ability to adapt its routing strategy to token-specific semantics. This observation aligns with our ablation study, which demonstrates that token-wise prediction yields better performance than sample-wise prediction.
- Since the router is jointly optimized with the generation tower, it can be regarded as a surrogate mechanism that approximates an optimal routing strategy. However, our analysis provides no evidence that the final-layer embedding serves as an effective solution. Similarly, we find no consistent pattern indicating a strict layer-to-layer correspondence in MoT. These findings suggest that previous designs may not fully leverage the capacity of the understanding tower, thereby supporting our hypothesis and underlying motivation.



(a) GenEval Score using different inference steps

(b) GenEval Score using different CFG scales.

(c) GenEval Score using or not using rescale strategy.

(d) GenEval Score using different schedulers.

Figure 6. **Ablation study results on GenEval [26] with different inference strategies.** MoS exhibits behavior similarly to other diffusion models. Incorporating commonly adopted enhancements into the inference process consistently improves performance.

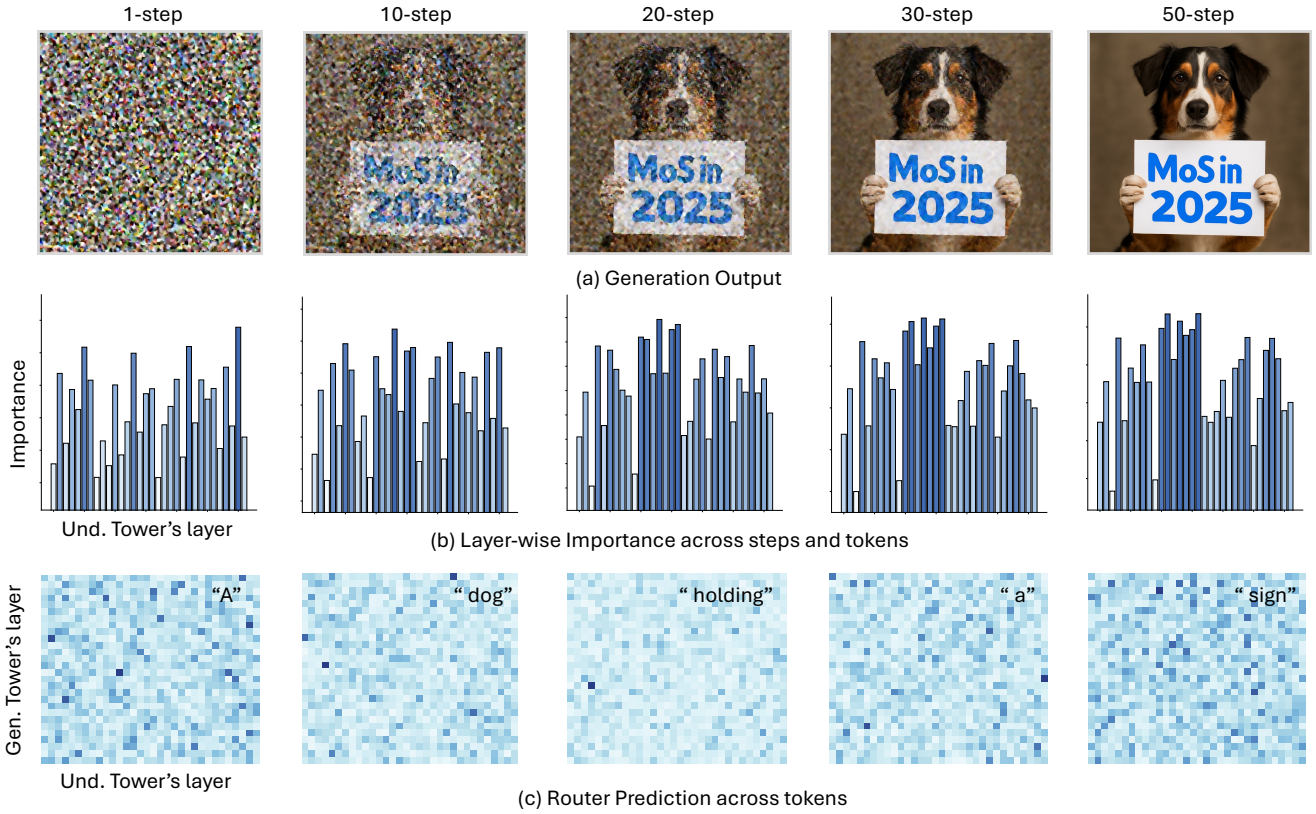


Figure 7. **Visualization of the Router across Time Steps.** The results show that different tokens induce distinct connection patterns, indicating that the router dynamically adjusts its layer-to-layer routing based on token-specific semantics.

4. Additional Benchmark Results

We provide comprehensive results across all benchmarks. Specifically, Table 2 reports the GenEval performance, Table 3 presents the DPG results, Table 4 shows the WISE score, and Table 5 lists the OneIG results. For image editing benchmarks, we report ImgEdit results in Table 6 and GEdit results in Table 7.

5. Additional Visualizations

We provide additional visualizations for clarity: i) Fig. 8 presents more text-to-image results generated by MoS; ii) Fig. 9 compares MoS with other advanced models under the same input captions; and iii) Fig. 10 illustrates comparisons with image editing models.

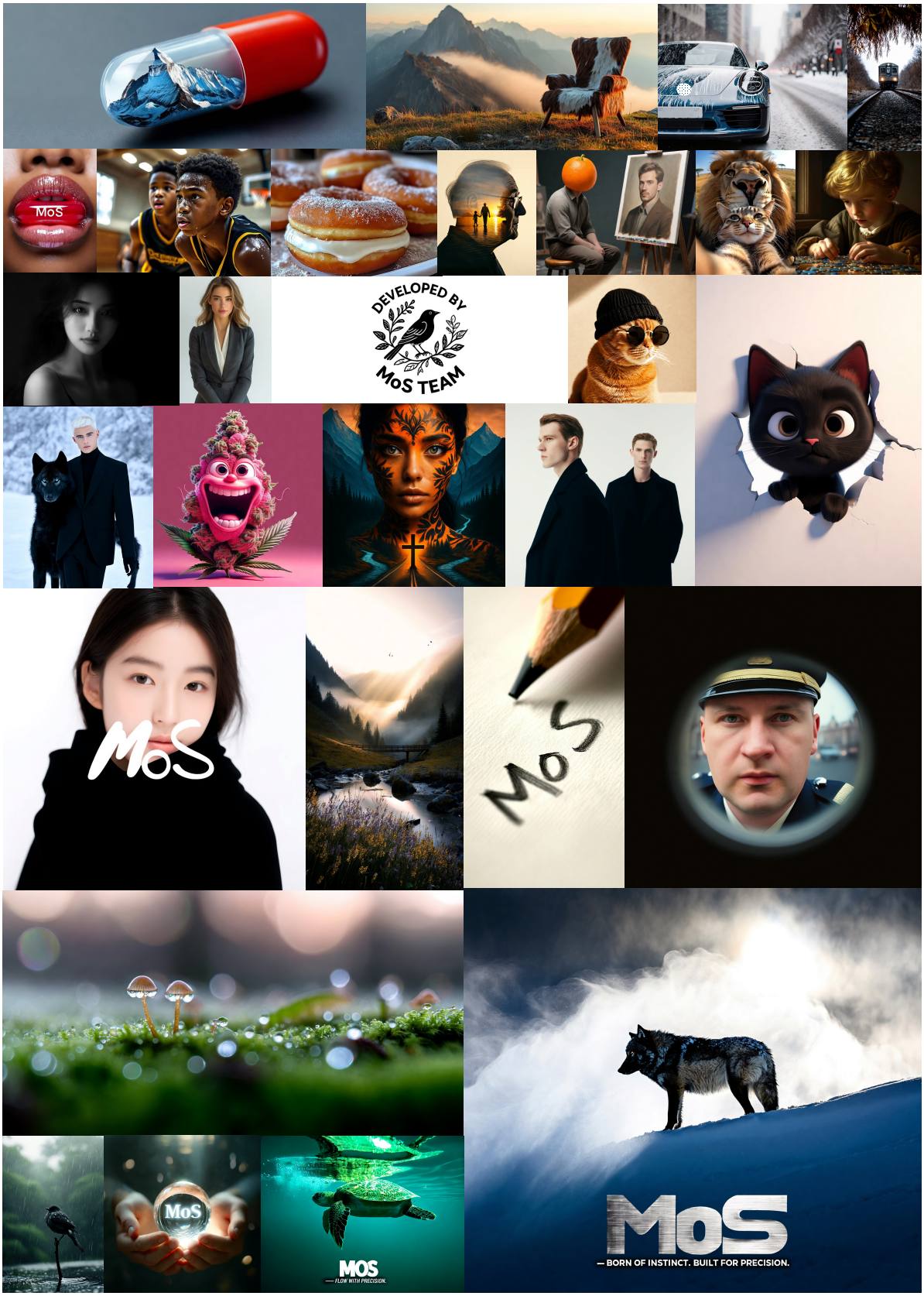


Figure 8. **Visualization of MoS-L on text-to-image generation.** The samples are produced under a dynamic resolution setting, with the maximum side length capped at 2048 pixels.

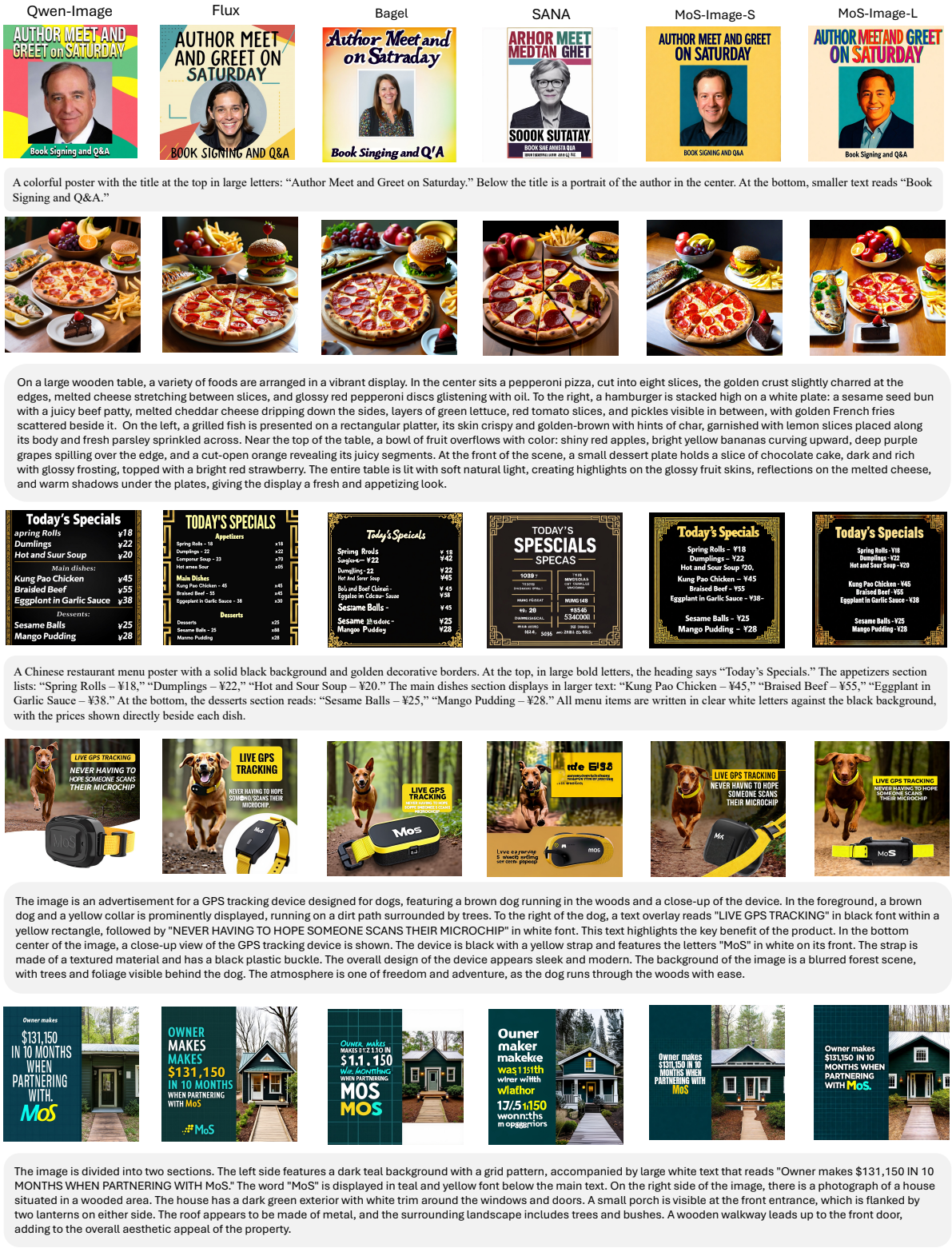


Figure 9. **Visualization of MoS-L/S and baseline methods on text-to-image generation.** All models are evaluated with their default parameters in `Diffusers`. We present results on challenging cases. These include scenarios such as arranging foods with distinct categories, colors, and patterns; posters combining natural objects with visual text; and purely textual prompts, such as generating a menu. MoS-L demonstrates competitive performance in these demanding settings. Zoomed-in view for better clarity.

Figure 10. **Visualization of MoS-L/S and baseline methods on instruction-based image editing.** All models are evaluated using their default parameters in `Diffusers`. We showcase results on hybrid instructions and the cases involving visual text editing. Zoomed-in for better clarity..

Table 2. **Performance of Foundational Image Generation Models on the GenEval Benchmark [26].** GenEval evaluates object-level prompt alignment in text-to-image models. Greyed rows denote models with unclear configuration references, which hinders fair comparison. MoS-L attains the strongest results across several dimensions and delivers the best overall performance. Here and below, #Param denotes the number of learnable parameters.

Model	#Param	Single Obj.↑	Two Obj.↑	Counting↑	Colors↑	Position↑	Color Attri.↑	Overall↑
GPT Image 1 [High] [59]	-	0.99	0.92	0.85	0.92	0.75	0.61	0.84
Seedream 3.0 [23]	-	0.99	0.96	**0.91**	**0.93**	0.47	0.80	0.84
Qwen-Image [79]	**20B**	0.99	0.92	**0.89**	0.88	0.76	0.77	0.87
Emu3-Gen [78]	8B	0.98	0.71	0.34	0.81	0.17	0.21	0.54
SD3 Medium [21]	2B	0.98	0.74	0.63	0.67	0.34	0.36	0.62
FLUX.1 [Dev] [43]	12B	0.98	0.81	0.74	0.79	0.22	0.45	0.66
SD3.5 Large [21]	8.1B	0.98	0.89	0.73	0.83	0.34	0.47	0.71
Lumina-Image 2.0 [65]	2.6B	-	0.87	0.67	-	-	0.62	0.73
Show-O2 [87]	7B	1.00	0.87	0.58	0.92	0.52	0.62	0.76
Janus-Pro [13]	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80
SANA-1.5 [85]	4.8B	0.99	0.93	0.86	0.84	0.59	0.65	0.81
HiDream-I1-Full [7]	17B	1.00	0.98	0.79	0.91	0.60	0.72	0.83
TAR [30]	7B	0.99	0.92	0.83	0.85	0.80	0.65	0.84
Bagel [18]	14B	0.98	0.95	0.84	**0.95**	0.78	0.77	0.88
Mogao [48]	7B	1.00	0.97	0.83	0.93	0.84	0.80	0.89
MoS-Image-S	3B	1.00	0.95	0.83	0.89	0.86	**0.81**	0.89
MoS-Image-L	5B	**1.00**	**0.97**	0.82	0.91	**0.88**	0.80	**0.90**

Table 3. **Performance of Foundational Image Generation Models on the DPG Benchmark [35].** DPG-Bench evaluates long-prompt alignment in text-to-image models. MoS-Image-L delivers state-of-the-art results across multiple dimensions, ranking just below Lumina.

Model	#Param	Global↑	Entity↑	Attribute↑	Relation↑	Other↑	Overall↑
DALL-E 3 [58]	-	90.97	89.61	88.39	90.58	89.83	83.50
GPT Image 1 [High] [59]	-	88.89	88.94	89.84	92.63	90.96	85.15
Seedream 3.0 [23]	-	**94.31**	**92.65**	91.36	92.78	88.24	88.27
Qwen-Image [79]	**20B**	91.32	91.56	92.02	94.31	**92.73**	**88.32**
SD v1.5 [68]	0.86B	74.63	74.23	75.39	73.49	67.81	63.18
SDXL [62]	6.6B	83.27	82.43	80.91	86.76	80.41	74.65
Playground v2.5 [45]	6.6B	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-DiT [46]	1.5B	84.59	80.59	88.01	74.36	86.41	78.87
PixArt-Σ [10]	0.6B	86.89	82.89	88.94	86.59	87.68	80.54
BLIP-3o [11]	8B	-	-	-	-	-	81.60
Emu3-Gen [78]	8B	85.21	86.68	86.84	90.22	83.15	80.60
FLUX.1 [Dev] [43]	12B	74.35	90.00	88.96	90.87	88.33	83.84
SD3 Medium [21]	2B	87.90	91.01	88.83	80.70	88.68	84.08
Janus-Pro [13]	7B	86.90	88.90	89.40	89.32	89.48	84.19
TAR [30]	7B	83.98	88.62	88.05	93.98	84.86	84.19
Mogao [48]	7B	82.37	90.03	88.26	93.18	85.40	84.33
HiDream-I1-Full [7]	17B	76.44	90.22	89.48	93.74	91.83	85.89
Show-o2-7B [87]	7B	89.00	91.78	89.96	91.81	91.64	86.14
Lumina-Image 2.0 [65]	2.6B	-	91.97	90.20	**94.85**	-	87.20
MoS-Image-S	3B	89.29	**92.17**	**92.09**	89.38	90.18	86.33
MoS-Image-L	5B	**91.74**	90.59	91.29	93.30	91.69	87.01

6. Algorithm Details

As shown in Algorithms 1–2, we provide the procedures for both training and inference of MoS.

7. Limitation and Future Studies

One-Way to Dual-Way Setting. MoT has demonstrated strong scalability under early-fusion training. In contrast, while MoS shows promising results for multimodal generation, its effectiveness in early-fusion settings remains to be validated. A principled extension is to endow the router with multiple projection layers to establish bidirectional transformer connections. We defer this exploration to future work due to computational and data constraints.

Human Preference Alignment. In this paper, we primarily adopt SFT as the post-training strategy for our models. Recent studies have explored applying CoT to multimodal generation [28] or employing GRPO to better align generated samples with human preferences [53]. Since our

Table 4. **Performance on world knowledge reasoning with WISE [57].** WISE evaluates complex semantic understanding and world knowledge in text-to-image generation.

Model	#Param	Cultural↑	Time↑	Space↑	Biology↑	Physics↑	Chemistry↑	Overall↑
GPT Image 1 [High] [59]	-	0.81	0.71	0.89	0.83	0.79	0.74	0.80
Qwen-Image [79]	**20B**	**0.62**	**0.63**	**0.77**	**0.57**	**0.75**	0.40	**0.62**
VILA-U [82]	7B	0.26	0.33	0.37	0.35	0.39	0.23	0.31
SDv1.5 [68]	0.86B	0.34	0.35	0.32	0.28	0.29	0.21	0.32
Janus-Pro [13]	7B	0.30	0.37	0.49	0.36	0.42	0.26	0.35
Emu3-Gen [78]	8B	0.34	0.45	0.48	0.41	0.45	0.27	0.39
SDXL [62]	6.6B	0.43	0.48	0.47	0.44	0.45	0.27	0.43
SD3.5 Large [21]	8.1B	0.44	0.50	0.58	0.44	0.52	0.31	0.46
PixArt-Alpha [10]	0.6B	0.45	0.50	0.48	0.49	0.56	0.34	0.47
Playground v2.5 [45]	6.6B	0.49	0.58	0.55	0.43	0.48	0.33	0.49
FLUX.1 [Dev] [43]	12B	0.48	0.58	0.62	0.42	0.51	0.35	0.50
BAGEL [18]	14B	0.44	0.55	0.68	0.44	0.60	0.39	0.52
UniWorld-V1 [49]	12B	0.53	0.55	0.73	0.45	0.59	0.41	0.55
MoS-Image-S	3B	0.40	0.50	0.65	0.43	0.63	0.37	0.47
MoS-Image-L	5B	0.47	0.56	0.74	0.49	0.64	**0.44**	0.54

Table 5. **Quantitative results on OneIG [8].** The overall score is averaged across five dimensions. With only 5B parameters, our model matches Imagen4 and trails recent commercial models by a small margin.

Model	# Param	Alignment↑	Text↑	Reasoning↑	Style↑	Diversity↑	Overall↑
Imagen3 [3]	-	0.84	0.34	0.31	0.36	0.19	0.41
Kolors 2.0 [76]	-	0.82	0.43	0.26	0.36	0.30	0.43
Recraft V3 [77]	-	0.81	0.80	0.32	0.38	0.21	0.50
Imagen4 [27]	-	0.86	0.81	0.34	0.38	0.20	0.52
Seedream 3.0 [23]	-	0.82	0.87	0.28	0.41	0.28	0.53
GPT Image 1 [High] [59]	-	0.85	0.86	**0.35**	**0.46**	0.15	0.53
Qwen-Image [79]	**20B**	**0.88**	**0.89**	**0.31**	**0.42**	0.20	**0.54**
Janus-Pro [13]	7B	0.55	0.00	0.14	0.28	0.37	0.27
BLIP3-o [11]	8B	0.71	0.01	0.22	0.36	0.23	0.31
BAGEL [18]	14B	0.77	0.24	0.17	0.37	0.25	0.36
Show-o2 [87]	7B	0.82	0.00	0.23	0.32	0.18	0.31
SDv1.5 [68]	0.86B	0.57	0.01	0.21	0.38	**0.43**	0.32
SDXL [62]	6.6B	0.69	0.03	0.24	0.33	0.30	0.32
SANA-1.5[85]	4.8B	0.77	0.07	0.22	0.40	0.22	0.33
Lumina-Image 2.0 [65]	2.6B	0.82	0.11	0.27	0.35	0.22	0.35
SD3.5 Large [21]	8.1B	0.81	0.63	0.29	0.35	0.23	0.46
FLUX.1 [Dev] [43]	12B	0.79	0.52	0.25	0.37	0.24	0.43
CogView4 [1]	6B	0.79	0.64	0.25	0.35	0.21	0.45
OmniGen2 [80]	4B	0.80	0.68	0.27	0.38	0.24	0.48
HiDream-I1-Full [7]	17B	0.83	0.71	0.32	0.35	0.19	0.48
MoS-Image-S	3B	0.82	0.82	0.26	0.38	0.20	0.50
MoS-Image-L	5B	0.85	0.87	0.26	0.41	0.19	0.52

model's behavior remains consistent with standard diffusion models, it can likewise benefit from such post-training techniques. We leave this direction for future work.

Efficiency Improvement. Our model is relatively smaller than prior state-of-the-art models, making it naturally more efficient. Nonetheless, it could be further accelerated through techniques such as low-precision quantization [84], model distillation [89], or feature caching [42, 52]. We leave these directions for future exploration.

Explainability. The MoS router predicts the relative importance of each potential connection, which offers a basis for interpreting cross-modal interactions. While this property may provide insights into model explainability, such analysis lies beyond the scope of this work and is left for future investigation.

Visual Artifacts. The primary goal of this paper is to address the challenge of instruction-following in multimodal generation. Nevertheless, our model still faces issues similar to other DiT or unified models, such as producing ar-

Table 6. **Performance of Foundational Image Editing Models on ImgEdit Benchmark [88].** Greyed rows indicate models lacking clear configuration references, which prevents fair comparison.

Model	#Param	Add↑	Adjust↑	Extract↑	Replace↑	Remove↑	Back.↑	Style↑	Hybrid↑	Action↑	Overall↑
FLUX.1 Kontext [Pro][4]	-	4.25	4.15	2.35	4.56	3.57	4.26	4.57	3.68	4.63	4.00
GPT Image 1 [High] [59]	-	4.61	4.33	2.90	4.35	3.66	4.57	**4.93**	3.96	**4.89**	4.20
Qwen-Image [79]	**20B**	4.38	4.16	**3.43**	4.66	4.14	4.38	4.81	3.82	4.69	4.27
MagicBrush [92]	0.86B	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
Instruct-Pix2Pix [6]	0.86B	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88
AnyEdit [90]	0.86B	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
UltraEdit [94]	2.5B	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.70
OmniGen [83]	3.8B	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
ICEdit [93]	0.2B	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
Step1X-Edit [55]	12B	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
BAGEL [18]	14B	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
UniWorld-V1 [49]	12B	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
OmniGen2 [80]	4B	3.57	3.06	1.77	3.74	3.20	3.57	**4.81**	2.52	**4.68**	3.44
MoS-Editing-S	3B	4.40	4.02	2.39	4.80	4.60	4.52	4.68	3.80	4.31	4.17
MoS-Editing-L	5B	**4.63**	**4.47**	2.04	**4.85**	**4.73**	**4.85**	4.71	**4.16**	4.52	**4.33**

Table 7. **Performance of Foundational Image Editing Models on GEdit Benchmark [55].** Greyed rows indicate models lacking clear configuration references, which prevents fair comparison.

Model	#param	G-Semantic Consistency↑	G-Perceptual Quality ↑	G.-Overall↑
Gemini 2.0 [17]	-	6.73	6.61	6.32
FLUX.1 Kontext [Pro][4]		7.02	7.60	6.56
GPT Image 1 [High] [59]	-	7.85	7.62	7.53
Qwen-Image [79]	20B	8.00	**7.86**	7.56
Instruct-Pix2Pix [6]	0.86B	3.58	5.49	3.68
AnyEdit [90]	0.86B	3.18	5.82	3.21
MagicBrush [92]	0.86B	4.68	5.66	4.52
UniWorld-V1 [49]	12B	4.93	7.43	4.85
OmniGen [83]	3.8B	5.96	5.89	5.06
OmniGen2 [80]	4B	7.16	6.77	6.41
BAGEL [18]	14B	7.36	6.83	6.52
Step1X-Edit [55]	12B	7.66	7.35	6.97
MoS-Editing-S	3B	8.00	7.34	7.41
MoS-Editing-L	5B	**8.54**	7.64	**7.86**

tifacts when the generated objects are very small (see our visualizations).

Algorithm 1 Training procedure of MoS

Require: Paired training data (z_0, c); understanding tower $\mathcal{U}$; generation tower $\mathcal{G}$; router $\mathcal{R}$; number of layers m, n; top-k_ϵ selection function.

1. Encode input.
Extract hidden states from the understanding tower:

$$\mathcal{U}(c) = \{\mathcal{S}_i^c \mid i \in [1, m]\}$$

2. Sample diffusion step.
Randomly sample timestep $t \sim \text{Uniform}(1, T)$ and obtain noisy latent z_t.
3. Predict routing weights.
Compute router logits:

$$\mathcal{W} = \mathcal{R}(c, t, z_t)$$

where $\mathcal{W} \in \mathbb{R}^{m \times n}$ and w_{ij} denotes the routing weight from the i-th understanding layer to the j-th generation layer.
Normalize logits:

$$\overline{\mathcal{W}} = \text{softmax}(\mathcal{W})$$

where the softmax operation is applied to each column $w_{1:m,j}$ of $\mathcal{W}$.
4. Construct conditional signal for each generation block.
for $j = 1$ to n **do**
    Select indices of top-k elements under ϵ-greedy rule:

$$I_j = \text{top-}k_\epsilon(\overline{w}_{1:m,j})$$

    Compute the conditional context:

$$\mathbf{S}_j^c = \sum_{i \in I_j} \overline{w}_{ij} \cdot \mathcal{S}_i^c.$$

    Fuse with the generation tower features:

$$\mathbf{H}_j = \text{Concat}\big(\text{Proj}(\mathbf{S}_j^c), \mathbf{S}_j^z\big)$$

    Perform Generation Tower Block-j on $\mathbf{H}_j$ to update $\mathbf{S}_j^z$ to $\mathbf{S}_{j+1}^z$.
end for
5. Compute loss.
Apply the diffusion objective (e.g., ℓ_2 loss) between predicted and ground-truth z_0.

Algorithm 2 Inference procedure of MoS

Require: Conditioning input c; understanding tower $\mathcal{U}$; generation tower $\mathcal{G}$; router $\mathcal{R}$; number of steps T; number of top connections k; number of layers m, n.
Ensure: Generated sample $\hat{z}_0$

1. Encode conditioning signal.
Obtain hidden states from the understanding tower:

$$\mathcal{U}(c) = \{\mathcal{S}_i^c \mid i \in [1, m]\}.$$

2. Initialize latent.
Sample initial noise $\hat{z}_1 \sim \mathcal{N}(0, \mathbf{I})$.
3. Iterative denoising.
for $t = 1$ down to 0 **do**
    Compute router logits:

$$\mathcal{W} = \mathcal{R}(c, t, \hat{z}_t)$$

    Normalize logits:

$$\overline{\mathcal{W}} = \text{softmax}(\mathcal{W})$$

    for $j = 1$ to n **do**
        Select top-k indices under ϵ-greedy rule:

$$I_j = \text{top-}k_\epsilon(\overline{w}_{1:m,j})$$

        Compute the conditional context:

$$\mathbf{S}_j^c = \sum_{i \in I_j} \overline{w}_{ij} \cdot \mathcal{S}_i^c.$$

        Fuse with the generation tower features:

$$\mathbf{H}_j = \text{Concat}\big(\text{Proj}(\mathbf{S}_j^c), \mathbf{S}_j^z\big)$$

        Perform Generation Tower Block-j on $\mathbf{H}_j$ to update $\mathbf{S}_j^z$ to $\mathbf{S}_{j+1}^z$.
    end for
    Predict $\hat{z}_{t-1}$ with $\hat{v}_t = \mathcal{G}(z_t, \mathcal{W}, \mathcal{U}(c))$ following the diffusion sampling rule.
end for
4. Decode.
Obtain final output $\hat{z}_0$ via the decoder of $\mathcal{G}$.

References

[1] Zhipu AI and THUDM / CogView Team. Cogview4. https://github.com/THUDM/CogView4, 2025. 11

[2] Sangmin Bae, Yujin Kim, Reza Bayat, Sungnyun Kim, Jiyoun Ha, Tal Schuster, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Aaron Courville, et al. Mixture-of-recursions: Learning dynamic recursive depths for adaptive token-level computation. *arXiv preprint arXiv:2507.10524*, 2025. 1

[3] Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluis Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024. 11

[4] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 12

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 1

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2023. 12

[7] Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025. 3, 10, 11

[8] Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *arXiv preprint arxiv:2506.07977*, 2025. 11

[9] Chen Chen, Rui Qian, Wenze Hu, Tsu-Jui Fu, Jialing Tong, Xinze Wang, Lezhi Li, Bowen Zhang, Alex Schwing, Wei Liu, et al. Dit-air: Revisiting the efficiency of diffusion model architecture design in text to image generation. *arXiv preprint arXiv:2503.10618*, 2025. 1

[10] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 74–91. Springer, 2024. 10, 11

[11] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. 1, 10, 11

[12] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Diffusion transformers for image and video generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[13] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 1, 10, 11

[14] Róbert Csordás, Kazuki Irie, Jürgen Schmidhuber, Christopher Potts, and Christopher D Manning. Moeut: Mixture-of-experts universal transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:28589–28614, 2024. 1

[15] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 1

[16] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 2

[17] Google DeepMind. Gemini 2.0. https://gemini.google.com/, 2025. 12

[18] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 1, 10, 11, 12

[19] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[20] David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013. 1

[21] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 1, 3, 10, 11

[22] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research (JMLR)*, 23(120):1–39, 2022. 1

[23] Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025. 10, 11

[24] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 1

[25] Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes

discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025. 1

[26] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:52132–52152, 2023. 4, 6, 10

[27] Google. Imagen. https://deepmind.google/models/imagen/, 2025. 11

[28] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Rui Huang, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Shanghang Zhang, Peng Gao, et al. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025. 10

[29] John Hampshire and Alex Waibel. Connectionist architectures for multi-speaker phoneme recognition. *Advances in neural information processing systems*, 2, 1989. 1

[30] Jiaming Han, Hao Chen, Yang Zhao, Hanyu Wang, Qi Zhao, Ziyan Yang, Hao He, Xiangyu Yue, and Lu Jiang. Vision as a dialect: Unifying visual understanding and generation via text-aligned representations. *arXiv preprint arXiv:2506.18898*, 2025. 10

[31] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 44(11):7436–7456, 2021. 1

[32] Alex Henry, Prudhvi Raj Dachapally, Shubham Pawar, and Yuxuan Chen. Query-key normalization for transformers. *arXiv preprint arXiv:2010.04245*, 2020. 1

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1

[34] Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, Siyu Zhu, Steven Shimizu, Shivam Sahni, Haowen Ning, Yanning Chen, and Zhipeng Wang. Liger-kernel: Efficient triton kernels for LLM training. In *Championing Open-source DEvelopment in ML Workshop @ ICML25*, 2025. 2

[35] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 10

[36] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 4

[37] Alexey Grigorevich Ivakhnenko. Polynomial theory of complex systems. *IEEE transactions on Systems, Man, and Cybernetics*, (4):364–378, 2007. 1

[38] Alekseĭ Grigor'evich Ivakhnenko and Valentin Grigor'evich Lapa. Cybernetic predicting devices. *Technical Report*, 1966.

[39] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 1

[40] Christopher Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E*, 1997. 1

[41] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 1

[42] Kumara Kahatapitiya, Haozhe Liu, Sen He, Ding Liu, Menglin Jia, Chenyang Zhang, Michael S Ryoo, and Tian Xie. Adaptive caching for faster video generation with diffusion transformers. *Proceedings of the International Conference on Computer Vision (ICCV)*, 2025. 1, 11

[43] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 10, 11

[44] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. 1

[45] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024. 10, 11

[46] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 10

[47] Weixin Liang, LILI YU, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Laerning Research (TMLR)*, 2025. 1, 3

[48] Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*, 2025. 1, 10

[49] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025. 11, 12

[50] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 5404–5411, 2024. 4

[51] Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*, 2024. 1

[52] Haozhe Liu, Wentian Zhang, Jinheng Xie, Francesco Faccio, Mengmeng Xu, Tao Xiang, Mike Zheng Shou, Juan-Manuel Perez-Rua, and Jürgen Schmidhuber. Faster diffusion through temporal attention decomposition. *Transactions on Machine Laerning Research (TMLR)*, 2025. 1, 5, 11

[53] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 10

[54] Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. Prismer: A vision-language model with an ensemble of experts. *arXiv preprint arXiv:2303.02506*, 2023. 1

[55] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 4, 5, 12

[56] Radford M Neal. Annealed importance sampling. *Statistics and computing*, 2001. 1

[57] Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Chaoran Feng, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025. 11

[58] OpenAI. Dall·e 3. https://openai.com/research/dall-e-3, 2025. Accessed: 2025-09-16. 10

[59] OpenAI. gpt-image-1. https://openai.com/index/image-generation-api/, 2025. OpenAI blog post introducing the model. Accessed: 2025-09-15. 10, 11, 12

[60] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 1

[61] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 1

[62] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 4, 10, 11

[63] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 4

[64] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 3

[65] Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, et al. Lumina-image 2.0: A unified and efficient image generative framework. *arXiv preprint arXiv:2503.21758*, 2025. 10, 11

[66] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[67] David Raposo, Sam Ritter, Blake Richards, Timothy Lillicrap, Peter Conway Humphreys, and Adam Santoro. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv preprint arXiv:2404.02258*, 2024. 1

[68] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 10, 11

[69] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[70] Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992. 1

[71] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 1

[72] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024. 1

[73] Qiao Sun, Zhicheng Jiang, Hanhong Zhao, and Kaiming He. Is noise conditioning necessary for denoising generative models? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025. 2

[74] Bingda Tang, Boyang Zheng, Sayak Paul, and Saining Xie. Exploring the deep fusion of large language models and diffusion transformers for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[75] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 1

[76] Kuaishou Kolors Team. Kolors2.0. https://app.klingai.com/, 2025. 11

[77] Recraft Team. Recraft v3. https://www.recraft.ai/, 2024. 11

[78] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 10, 11

[79] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 1, 3, 10, 11, 12

[80] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie

Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 11, 12

[81] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 1

[82] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 11

[83] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13294–13304, 2025. 12

[84] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. 11

[85] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025. 10, 11

[86] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. 1

[87] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 1, 10, 11

[88] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025. 12

[89] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6613–6623, 2024. 11

[90] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26125–26135, 2025. 12

[91] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 1

[92] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:31428–31449, 2023. 12

[93] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025. 12

[94] Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:3058–3093, 2024. 12

[95] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 1