TAG-MoE: Task-Aware Gating for Unified Generative Mixture-of-Experts

Yu Xu[1,2†] Hongbin Yan[1] Juan Cao[1] Yiji Cheng[2] Tiankai Hang[2]
Runze He[2] Zijin Yin[2] Shiyi Zhang[2] Yuxin Zhang[1] Jintao Li[1]
Chunyu Wang[2‡] Qinglin Lu[2] Tong-Yee Lee[3] Fan Tang[1§]
[1]University of Chinese Academy of Sciences [2]Tencent Hunyuan [3]National Cheng-Kung University

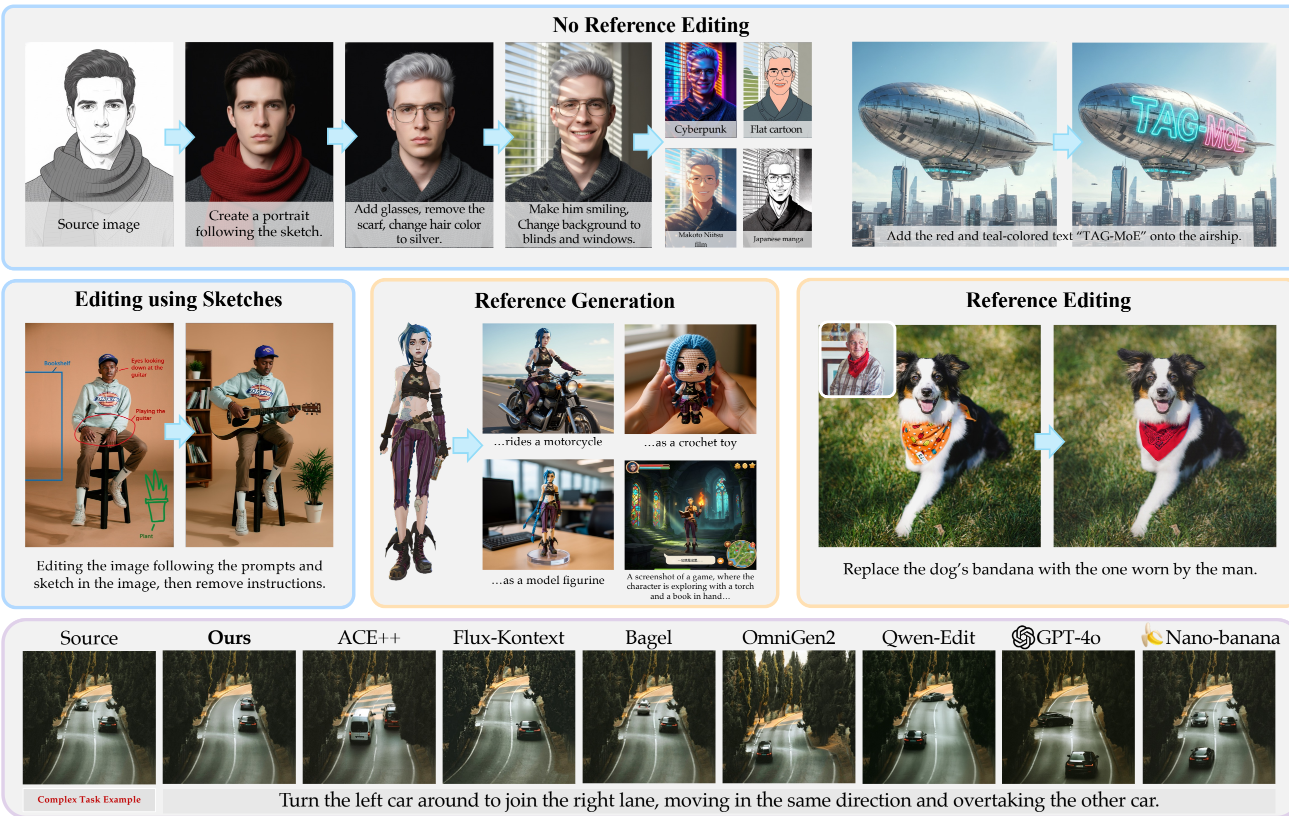

Figure 1. We present **TAG-MoE**, by injecting high-level task semantic intent into the local routing decisions of the MoE gating network, we enabling the diffusion transformer model to handle diverse generative tasks.

Abstract

Unified image generation and editing models suffer from severe task interference in dense diffusion transformers architectures, where a shared parameter space must compromise between conflicting objectives (e.g., local editing v.s. subject-driven generation). While the sparse Mixture-of-Experts (MoE) paradigm is a promising solution, its gating networks remain task-agnostic, operating based on local features, unaware of global task intent. This task-agnostic nature prevents meaningful specialization and fails to resolve the underlying task interference. In this paper, we propose a novel framework to inject semantic intent into MoE routing. We introduce a Hierarchical Task Semantic Annotation scheme to create structured task descriptors (e.g., scope, type, preservation). We then design Predictive Alignment Regularization to align internal routing decisions with the task's high-level semantics. This regularization evolves the gating network from a task-agnostic executor to a dispatch center. Our model effectively mitigates task interference, outperforming dense baselines in fidelity and quality, and our analysis shows that experts naturally develop clear and semantically correlated specializations.

† Work done during internship at Tencent Hunyuan.
‡ Project leader.
§ Corresponding author. tfan.108@gmail.com

1. Introduction

The field of visual synthesis is rapidly converging toward unified image generation and editing models [7, 20, 21, 49], frameworks designed to consolidate disparate image manipulation tasks—from subject customization and style transfer to high-fidelity inpainting and instruction-based editing—into a single, robust system with the help of large-scale, dense Diffusion Transformers (DiT).

While promising efficiency, this unification is critically bottlenecked by severe task interference. The shared parameter space must simultaneously execute inherently contradictory objectives: local editing demands precise content preservation, while subject-driven generation requires expressive diversity and novel synthesis. This fundamental conflict forces the network toward a "mediocre compromise solution," preventing the necessary representational specialization and ultimately degrading performance across the spectrum of user intents.

To overcome the scalability [13] and capacity [56] limits of dense DiT, the sparse Mixture-of-Experts (MoE) paradigm is adopted to dramatically expand model capacity with manageable inference costs of large-scale generative models. However, these efforts mainly focus on single, general-purpose image generation tasks, and have not (and do not need to) account for the complex task diversity within the unified generation framework. Applying standard MoE to the heterogeneous unified domain introduces a critical architectural failure: the **task-agnostic** nature of conventional gating networks. Standard routers rely solely on local token features, remaining entirely oblivious to the high-level, global task intent (e.g., "identity preservation" or "style modification"). This profound information gap between the local gate and the global objective leads to spontaneous, inefficient expert specialization, fundamentally failing to structurally disentangle multi-task interference. *How to inject the high-level, global task semantics into the local MoE routing mechanism to enable task-aware specialization remains an open challenge.*

In this study, we propose **TAG-MoE**, a task-aware gating network for unified image generation and editing. First, to provide a structured unified task representation, we introduce a **hierarchical task semantic annotation** scheme, by decomposing specific generative task into a multi-faceted descriptor, capturing the operational *scope* (e.g., local/-global editing), the semantic *type* (e.g., attribute/action editing), and essential *preservation constraints* (e.g., identity/style preservation). Such structured representations provides the necessary rich supervisory signal previously missing. Furthermore, we propose a novel training framework founded on the principle that semantically similar generation tasks evokes similar expert usage patterns. To enforce this, we design an innovative **predictive alignment regularization** to correlate the high-level task semantic intent with the underlying routing decisions. Such regularization serves as a bridge to compel the model's internal routing strategy to become predictive of the task's macro-semantics, injecting global semantic intent into the local routing mechanism, leading the gating network to evolve from a task-agnostic executor into an aware, intelligent dispatch center. Experiments on unified image generation benchmarks ICE-Bench, image editing benchmark EmuEdit and GEdit, subject-driven generation benchmark DreamBench++ and OmniContext indicate that our method achieves the best overall performance. Our primary contributions are summarized as follows:

1. We propose a novel task-aware sparse MoE framework and successfully apply it to Diffusion Transformer-based unified image generation and editing tasks.
2. We introduce a hierarchical task semantic annotation scheme and a corresponding predictive alignment regularization that, together, effectively resolve the task-agnostic of the MoE gate by aligning its routing strategy with the task's semantic intent.
3. By successfully mitigating task interference, our model achieves SOTA overall performance against open-source baselines across five comprehensive benchmarks.

2. Related Work

2.1. Unified Image Generation and Editing

Recent efforts in unified image generation aim to build single models capable of handling a broad range of image manipulation tasks, moving beyond specialized, task-specific approaches. Early methods treat the problem as a sequence-to-sequence task, concatenating text, source, and target image tokens for large transformers [11, 12, 14, 16, 45]. Subsequent works refine input representations and architectures to improve multimodal conditioning. Methods such as UniReal [5], RealGeneral [23] and X2Edit [25] introduces trainable index, subject, condition and task embeddings to enhance alignment, while Flux-Kontext [21] employs 3D rotary positional encodings to distinguish source from target images. Architectural innovations include dual-branch models that decouple subject and background processing [22], channel-wise concatenation to preserve contextual signals [26], and the integration of auxiliary MLLMs or transformers for improved scene understanding [10, 19, 36, 42, 50], albeit with increased complexity and compute.

Despite these advances, current unified models overlook a central challenge: the inherent conflict between the objectives of different image-to-image tasks. Editing tasks [2, 17, 18, 46, 57] (e.g., style transfer, object removal) require precise regional preservation while modifying others, whereas customization tasks [33, 47, 48, 53] (e.g., subject-driven generation) demand strong identity consistency across new contexts. Without explicitly modeling these distinct—and often competing—requirements, existing approaches struggle to adaptively serve the full spectrum of user intents, limiting their practical robustness and generalization.

2.2. Image Generation with Mixture of Experts

The MoE paradigm increases model capacity by routing inputs to specialized sub-networks, or "experts," avoiding a proportional rise in per-sample computation. Its success in large language models has motivated adoption in visual generation: pioneering works such as DiT-MoE [13], and scaled variants like HunyuanImage-3.0 [3] and Dense2MoE [56], show that sparse expert architectures can enhance the expressiveness of diffusion transformers. Extending MoE to image editing, ICEdit [54] integrates LoRA-based MoE modules into attention blocks. However, purely data-driven routing is fundamentally limited: task-agnostic routers cannot resolve conflicts between heterogeneous tasks (e.g., editing vs. customization), and the restricted capacity of LoRA experts hampers learning multi-task behaviors. Our approach overcomes these limitations by introducing task-aware expert routing. We condition the gating mechanism on learnable embeddings corresponding to specific task categories, enabling dynamic selection of the most relevant experts. This mitigates inter-task conflicts, promotes effective specialization, and achieves superior performance across diverse image-to-image tasks while maintaining the efficiency of the MoE framework.

3. Method

Our unified framework (Fig. 2) employs a Multimodal Diffusion Transformer (MM-DiT) with MoE layers for efficient, dynamic task handling (§3.1). We introduce hierarchical task semantic annotation (§3.2) and a novel semantic-aligned router (§3.3). This router guides the MoE's specialization by aligning its routing decisions with these explicit task semantics in an interpretable manner .

3.1. MoE-based Multimodal Diffusion Transformer

Building upon an MM-DiT architecture, our approach processes diverse inputs within a unified token sequence framework. To interpret user instructions, we employ a powerful pre-trained Multimodal Large Language Model (MLLM) to encode the input text c_{text} into a sequence of text embeddings C. Separately, a pre-trained VAE encoder $\mathcal{E}$ maps both the conditional image I_c and the target image I_0 into latent representations, z_c and z_0. During training, Gaussian noise is sampled and added to the z_0 to produce a noisy version z_t. Both z_c and z_t are then patchified into sequences of visual tokens. Finally, the complete input to our MM-DiT is a single sequence formed by concatenating the text embeddings C, the image tokens from z_c, the image tokens from the noisy target latent z_t, and a timestep embedding [30].

We replace the feed-forward networks (FFNs) of the image stream in diffusion transformer blocks with MoE layers. This leverages sparse activation to significantly increase model capacity at a fixed activation parameter, enabling superior performance over dense models with a comparable budget. We only implement MoE layers in the later transformer blocks as high-level semantic synthesis in these deeper layers benefits most from the increased capacity [9, 32]. The MoE layer consists of a set of N expert networks E and a gating network $\mathcal{G}$. The gating network $\mathcal{G}$ maps each input token to a probability distribution over the N experts, thereby determining their top k selections $\mathcal{T} \subseteq \{E_1, \ldots, E_N\}$. The output is a weighted sum of the activated experts' outputs:

$$\text{MoE}(x) = \sum_{E_i \in \mathcal{T}(x)} \mathcal{G}(x)_i \cdot E_i(x). \tag{1}$$

This MoE-enhanced architecture is trained end-to-end using a Flow Matching objective.

3.2. Hierarchical Task Semantic Annotation

To train a unified model that supports a broad range of generation and editing tasks, a structured representation of task semantics is essential. A single coarse label (e.g., "edit") cannot capture user intent. For example, "change the background to a beach" and "make the person smile' are both edits but require fundamentally different behaviors and preservation constraints. To address this, we introduce a three-tier annotation scheme that provides each training instance (source image, instruction, target image) with a rich semantic descriptor: Scope - the task's operational nature and spatial extent (e.g., global editing, local editing, content customization). Type — the semantic category of the manipulation (e.g., object editing, style transfer, attribute editing). Preservation — the invariants that must remain unchanged (e.g., identity, background, structure preservation).

An automated pipeline utilizing Qwen-VL [1] is established to analyze training triplets. It involves providing definitions of a three-tier system and instructing Qwen-VL to output atomic tags. The rule set is continuously refined to maintain consistency and semantic quality.

For instance, the task "Make the person in the photo wear sunglasses" would be annotated with tags such as "Scope: local editing; Type: object editing; Preservation: identity preservation, background preservation, style preservation". This rich set of atomic tags forms the basis for our semantic representation. A comprehensive list of all tags and the corresponding rule set are provided in the supplementary materials.

Inference Stage. This hierarchical annotation scheme is exclusively used for training. During the inference stage, these ground-truth tags are no longer required. Instead, as

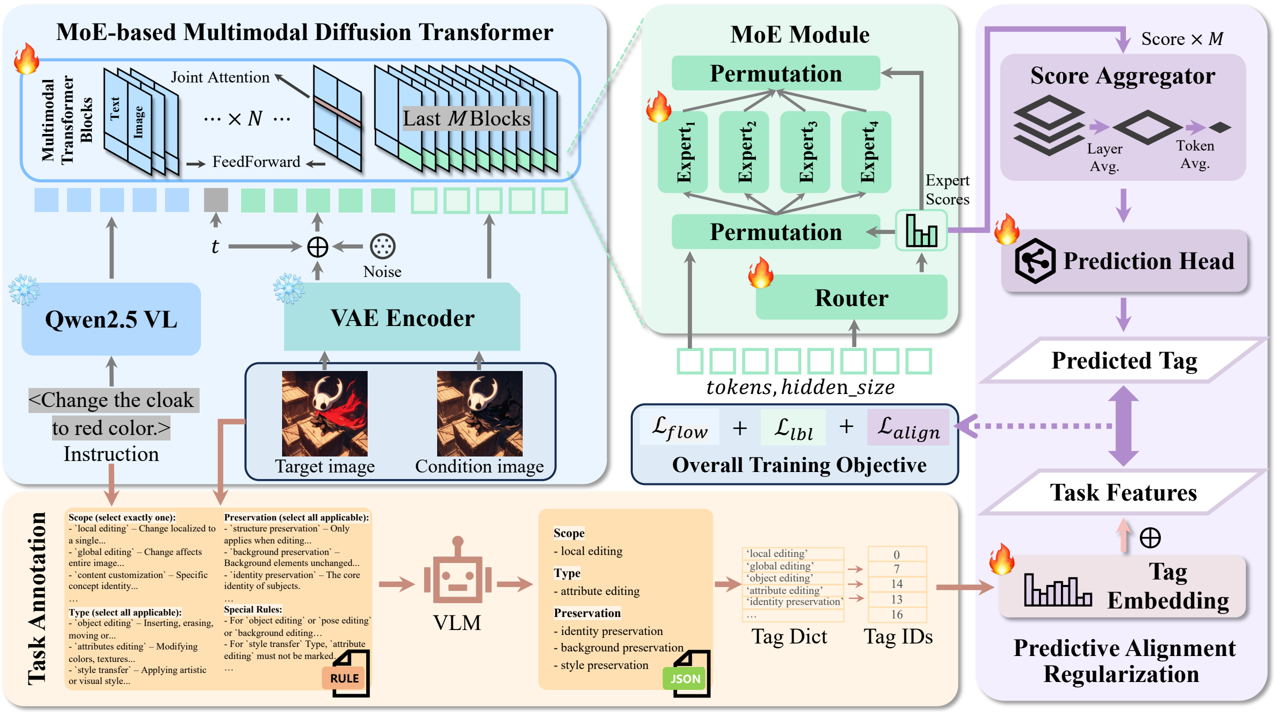

Figure 2. Pipeline of our method. TAG-MoE consists of: (1) A **MM-DiT with MoE** layers; (2) A **Hierarchical Task Semantic Annotation** that labels training data with atomic task descriptors; (3) A novel Semantic-Aligned Router explicitly aligns MoE routing behavior with task semantics through **Predictive Alignment Regularization**.

a lightweight pre-processing step, we pass the user's raw instruction c_{text} and the source image I_c to a VLM (e.g., Qwen-VL [1]). The VLM performs instruction rewriting, analyzing the image and text to generate a more detailed, descriptive prompt. This enriched prompt is then encoded as the text embedding C and fed into the MM-DiT.

3.3. Semantic-Aligned Gating Network

We design a novel semantic-aligned gating network to force the model's internal routing strategy (encoded as a routing signature "$\mathbf{g}$") to predict the task's macroscopic semantics (encoded as a semantic embedding "$\mathbf{s}$"). This predictive alignment serves as a bridge, connecting local routing decisions with global task intent. Our mechanism comprises three key components: (1) construction of the global semantic embedding $\mathbf{s}$; (2) construction of the aggregated routing signature $\mathbf{g}$; and (3) the predictive alignment loss $\mathcal{L}_{align}$.

3.3.1. Global Semantic Embedding

Based on the hierarchical task semantic annotation described in §3.2, we first define a global vocabulary $\mathcal{V}$ containing all K atomic tags (e.g., "local editing", "identity preservation"). We instantiate a learnable tag embedding matrix $\mathbf{W}_{tag} \in \mathbb{R}^{K \times D}$ for this vocabulary, where D is the model's hidden dimension. For a given training sample, its associated tags form a set $T_p \subseteq \mathcal{V}$ (e.g., $T_p = \{$"local editing", "face preservation"$\}$). To convert this variable-sized set T_p into a fixed-dimension vector $\mathbf{s}$, we first retrieve the corresponding embedding vector $\mathbf{e}_t = \mathbf{W}_{tag}[\text{index}(t)]$ for each tag $t \in T_p$, and then aggregate them via element-wise summation. This constructs the global semantic embedding $\mathbf{s}$, which represents the "macro-level semantic ground truth":

$$\mathbf{s} = \sum_{t \in T_p} \mathbf{W}_{tag}[\text{index}(t)]. \tag{2}$$

This vector $\mathbf{s} \in \mathbb{R}^D$ is permutation-invariant, meaning the order of tags does not affect the final representation. It serves as the structured supervisory signal for our subsequent alignment loss.

3.3.2. Aggregated Routing Signature

Correspondingly, we require a vector to represent the internal routing strategy the model actually employs for the current sample. The gating network $\mathcal{G}$ (see §3.1) generates routing scores $S_{l,t} \in \mathbb{R}^N$ for each token t in each of the L MoE layers, where N is the number of experts.

To obtain a single vector representing the expert usage pattern for the entire sample, we design an aggregated routing signature $\mathbf{g}$. First, we average the routing scores across all L MoE layers to get a per-token average score $\bar{S}_t = \frac{1}{L}\sum_{l=1}^{L} S_{l,t}$. Next, we apply mean pooling over the sequence (token) dimension to get the final signature $\mathbf{g} \in \mathbb{R}^N$:

$$\mathbf{g} = \frac{1}{T}\sum_{t=1}^{T} \bar{S}_t = \frac{1}{T \cdot L}\sum_{t=1}^{T}\sum_{l=1}^{L} S_{l,t}. \tag{3}$$

This vector $\mathbf{g}$ encodes which experts are activated on average to process the sample, capturing its *de facto* internal routing policy.

3.3.3. Predictive Alignment Regularization

We now have two vectors: $\mathbf{s} \in \mathbb{R}^D$, representing what the task *should be*, and $\mathbf{g} \in \mathbb{R}^N$, representing what the model *actually do*. To align them, we introduce a lightweight prediction head $\mathcal{H}_{pred}$ (a two-layer MLP), to project the aggregated routing signature $\mathbf{g}$ from the expert space $\mathbb{R}^N$ into the semantic space $\mathbb{R}^D$, yielding a predicted semantic embedding $\hat{\mathbf{s}} = \mathcal{H}_{pred}(\mathbf{g})$.

We force the routing strategy to predict the task semantics by minimizing the cosine similarity loss between $\hat{\mathbf{s}}$ and $\mathbf{s}$. This is our Predictive Alignment Loss $\mathcal{L}_{align}$:

$$\mathcal{L}_{align} = 1 - \text{sim}(\hat{\mathbf{s}}, \mathbf{s}) = 1 - \frac{\hat{\mathbf{s}} \cdot \mathbf{s}}{|\hat{\mathbf{s}}||\mathbf{s}|}. \tag{4}$$

Minimizing $\mathcal{L}_{align}$ trains the parameters of $\mathcal{H}_{pred}$ and, more importantly, backpropagates the gradient through $\mathbf{g}$ to the gating networks $\mathcal{G}$ of all MoE layers. This compels $\mathcal{G}$ to evolve from a task-agnostic executor into a semantic-aware scheduler: it must learn to route tokens intelligently, such that the resulting aggregate signature $\mathbf{g}$ contains sufficient information to predict the global task $\mathbf{s}$.

3.3.4. Overall Training Objective

Our proposed $\mathcal{L}_{align}$ is an auxiliary loss that complements the model's primary objective. The final overall loss $\mathcal{L}_{total}$ is a weighted sum of the main generation loss (e.g., $\mathcal{L}_{flow}$), the standard MoE load balancing loss $\mathcal{L}_{lbl}$, and our semantic alignment loss $\mathcal{L}_{align}$:

$$\mathcal{L}_{total} = \mathcal{L}_{flow} + \lambda_{lbl}\mathcal{L}_{lbl} + \lambda_{align}\mathcal{L}_{align}, \tag{5}$$

where λ_{lbl} and λ_{align} are hyperparameters that balance the contribution of each loss term.

3.4. Dataset Construction

Our model is trained on a large-scale, diverse dataset comprising both publicly available and proprietary in-house data, totaling over 11 million samples. This hybrid approach ensures broad coverage across the unified task space. The public portion (2.2M samples) is compiled from established benchmarks, including InstructP2P [2], UltraEdit [55], and OmniEdit [40] for universal instructive editing, supplemented by VTON-HD [6] for virtual try-on tasks and Ominicontrol [37] for subject driven generation.

Our proprietary in-house dataset is meticulously constructed using a multi-stage pipeline to cover a wide spectrum of specialized tasks. First, we source pristine images from large-scale public datasets. Next, we employ large language models (e.g., GPT-4o [28]) to generate a vast array of diverse editing and generation instructions for these images. To obtain high-quality target images, we utilize a combination of specialist and generalist models: for instance, specialist models like ControlNet [52] are used for "Control generation" tasks, while powerful generalist models (e.g., Flux-Kontext [21], Qwen-Edit [41], and SeedEdit [39]) are employed for a broad range of edits. Following the methodology of UniReal [5], we also process video frames to create dynamic editing datasets (e.g., for pose/view changes). Finally, to enhance robustness and quality, we systematically augment the data by constructing corresponding inverse tasks and instructions (e.g., pairing "object addition" with "object removal"), which significantly improves generative fidelity. **Detailed statistics of the dataset are provided in the supplementary material.**

4. Experiments

4.1. Implenentation Details

Our model is based on Qwen-Image T2I model [41], we integrate the MoE layers by replacing the standard FFNs of the image stream in the final 10 layers of our diffusion transformer. Each MoE layer consists of four experts, where each expert possesses an architecture identical to the original FFN it replaces. The gating network is implemented as a two-layer MLP, and we employ a top-1 routing strategy.

4.2. Experiments Settings

Baselines. We compare our method against three categories of SOTA baselines. (1) Unified generation and editing methods for diverse image-to-image tasks, including ACE++ [26], Flux.1 Kontext [21], BAGEL [10], OmniGen2 [42], Qwen-Edit [42] and DreamOmni2 [44]. We also include comparisons against product-level, closed-source models (e.g. GPT-4o [28] and Gemini-2.5-flash (aka. Nanobanana) [15], to contextualize our performance. However, our primary quantitative evaluation and main claims are benchmarked against open-source baselines. (2) Specialized zero-shot instruction-based editing methods, including InstructPix2Pix [2], EmuEdit [34], MagicBrush [51], UltraEdit [55], ICEdit [29], and Step1X-Edit [24]. (3) Specialized zero-shot subject-driven generation methods, including DreamO [27], OminiControl [37] and UNO [43].

Evaluation benchmarks. To comprehensively assess our model in the unified image generation and editing setting, we adopt ICE-Bench [29] as our primary benchmark, as it is specifically designed for unified models and spans both diverse editing tasks and subject-driven generation. For more fine-grained evaluation, we further include specialized benchmarks: EmuEdit-Bench [34] and GEdit-Bench [24] for detailed editing analysis, and DreamBench++ [31] together with OmniContext [42] to evaluate subject-driven generation performance.

Method	Aes.	CLIP-src	CLIP-cap	CLIP-ref	vllmqa
ACE++	5.219	0.851	0.263	0.713	0.637
Kontext	5.165	0.863	0.274	0.728	0.629
BAGEL	4.757	0.863	0.276	0.687	0.699
OmniGen2	5.238	0.855	0.279	0.728	0.787
Qwen-Edit	5.358	0.840	0.279	0.671	0.774
DreamOmni2	5.188	**0.866**	0.268	**0.739**	0.664
Ours	**5.399**	0.857	**0.282**	0.732	**0.852**
GPT-4o	5.801	0.823	0.278	0.693	0.889
Gemini-2.5-flash	5.571	0.879	0.281	0.724	0.847

Table 1. Comparison results for unified tasks on ICE-Bench [29] test sets. Open-source models are in the first block and close-source produce-level models are in the second block.

Metrics. We employ a comprehensive set of metrics to evaluate both visual quality and task correctness. Aesthetic quality is assessed using a SigLip-based predictor. Consistency with the source image is measured via CLIP-src (for editing) and CLIP-ref (for subject-driven generation), while text alignment is captured by CLIP-cap. For editing evaluation, we further use Qwen2-VL-72B [38] to determine whether the instruction is correctly executed based on the source image, instruction, and output image, yielding the vllmqa score. For subject-driven tasks, we assess three key preservation dimensions: facial identity (Face-ref, using the buffalo model from InsightFace App [8]), subject similarity (DINO-ref, via DINO [4]), and style fidelity (Style-ref, via CSD [35]). All metrics not originally within the [-1, 1] range are normalized. For every metric reported, higher values indicate better performance. In the tables, the best results are highlighted **in bold**, and the second-best results are underlined.

4.3. Quantitative Comparison

Unified generation evaluation. We report the main results on ICE-Bench in Tab. 1. Our method achieves the highest scores among all open-source baselines across three key metrics: aesthetic quality, CLIP-cap, and vllmqa. Notably, our CLIP-cap score not only surpasses all open-source competitors but also exceeds closed-source, product-level models such as GPT-4o and Gemini-2.5-flash, indicating stronger alignment with user instructions across diverse generation and editing tasks. Although some baselines exhibit high source fidelity (e.g., DreamOmni2 on CLIP-src), our model attains a more favorable overall balance by excelling in instruction adherence and semantic alignment.

We further present a per-category breakdown over 26 task types on ICE-Bench, visualized in the radar charts in Fig. 4. Our model achieves state-of-the-art performance in the vast majority of categories, demonstrating robust and well-balanced capability. DreamOmni2's high reference-generation scores largely stem from copy-paste behavior on source subjects, which artificially inflates similarity metrics.

Image editing evaluation We further evaluate our model against specialized zero-shot editing baselines on EmuEdit-bench [34] and GEdit-bench [24], with results shown in Tab. 2. (Note: Since EmuEdit is not open-source and only provides pre-generated outputs on its own benchmark, its performance on GEdit-bench is unavailable.) Although our model does not achieve top-1 performance on every metric, it clearly leads on the most important indicator vllmqa achieving the highest scores on both benchmarks. This is particularly noteworthy because, unlike static CLIP similarity, vllmqa uses a powerful VLLM to evaluate the correctness of the executed instruction, offering a more intelligent and reliable measure of editing success. Our strong results on this metric underscore the model's advanced instruction-following capability.

Method	EmuEdit-bench			GEdit-bench		
	CLIP-src	CLIP-cap	vllmqa	CLIP-src	CLIP-cap	vllmqa
InsP2P	0.8589	0.2919	0.2507	0.8604	0.3192	0.3191
EmuEdit	0.8854	0.3098	0.6253	-	-	-
MagicBrush	0.8552	0.2951	0.4573	0.8068	0.3146	0.3783
UltraEdit	0.8625	0.3075	0.3609	0.8459	0.3323	0.4605
ICEdit	0.8912	0.3026	0.3609	0.9007	0.3283	0.4145
Step1X-Edit	0.8845	0.3119	0.7893	0.8967	0.346	0.8158
ACE++	0.8367	0.2385	0.0606	0.8160	0.2518	0.0559
Kontext	**0.9091**	0.3093	0.741	0.9190	0.3419	0.7303
BAGEL	0.8565	0.3129	0.7989	0.8727	0.3470	0.7961
OmniGen2	0.8932	0.3087	0.5978	0.8940	0.3373	0.6546
Qwen-Edit	0.8832	**0.3159**	0.9174	0.9104	**0.3522**	0.875
DreamOmni2	0.9035	0.3096	0.6997	0.9229	0.3401	0.6349
Ours	0.9054	0.3152	**0.9284**	**0.9238**	0.3485	**0.8854**

Table 2. Comparison of instruction-based editing methods on EmuEdit-bench and GEdit-bench with multiple metrics.

Method	DreamBench++					OmniContext				
	CLIP-cap	CLIP-ref	DINO-ref	Face-ref	Style-ref	CLIP-cap	CLIP-ref	DINO-ref	Face-ref	Style-ref
DreamO	0.2899	0.7792	0.7518	0.335	0.5355	0.2986	0.7302	0.7075	0.4522	-
Ominicontrol	0.296	0.7642	0.6991	0.0579	0.3876	0.3067	0.7009	0.6126	-	-
UNO	0.2832	0.776	0.7429	0.2572	0.4328	0.2962	0.7106	0.6961	0.3665	-
ACE++	0.2791	0.7759	0.732	0.1636	0.5306	0.2832	0.7183	0.6932	0.1789	-
Kontext	0.2829	**0.819**	0.7919	0.3429	0.5655	0.2962	**0.765**	0.7494	0.5596	-
BAGEL	**0.3036**	0.7338	0.6998	0.0487	0.5065	0.2914	0.7188	0.7094	0.1264	-
OmniGen2	0.298	0.7712	0.752	0.1213	0.5167	0.3056	0.7544	0.7289	0.3919	-
Qwen-Edit	0.3009	0.7595	0.7187	0.2188	0.5095	**0.3152**	0.7115	0.6797	0.3019	-
DreamOmni2	0.2731	0.8062	**0.8008**	0.2344	0.5364	0.2848	0.7733	0.7611	0.5111	-
Ours	0.3011	0.7906	0.7613	**0.3678**	**0.5679**	0.3096	0.7297	**0.7628**	**0.5607**	-

Table 3. Comparison of subject-driven generation methods on DreamBench++ and OmniContext with multiple metrics.

Subject driven evaluation. We evaluate our model's fine-grained preservation ability against specialized subject-driven generation methods on DreamBench++ and OmniContext, with results shown in Tab. 3. We focus on metrics that measure subject, identity, and style fidelity (noting that OmniContext does not include style-related tasks). The results indicate strong preservation performance: our model achieves SOTA Face-ref scores on both benchmarks and the highest Style-ref score on DreamBench++. In addition, we obtain the top DINO-ref score on OmniContext and remain highly competitive on DreamBench++. These findings demonstrate that our unified model can match or surpass specialized models, effectively mitigating the typical tension between subject fidelity and generative diversity.

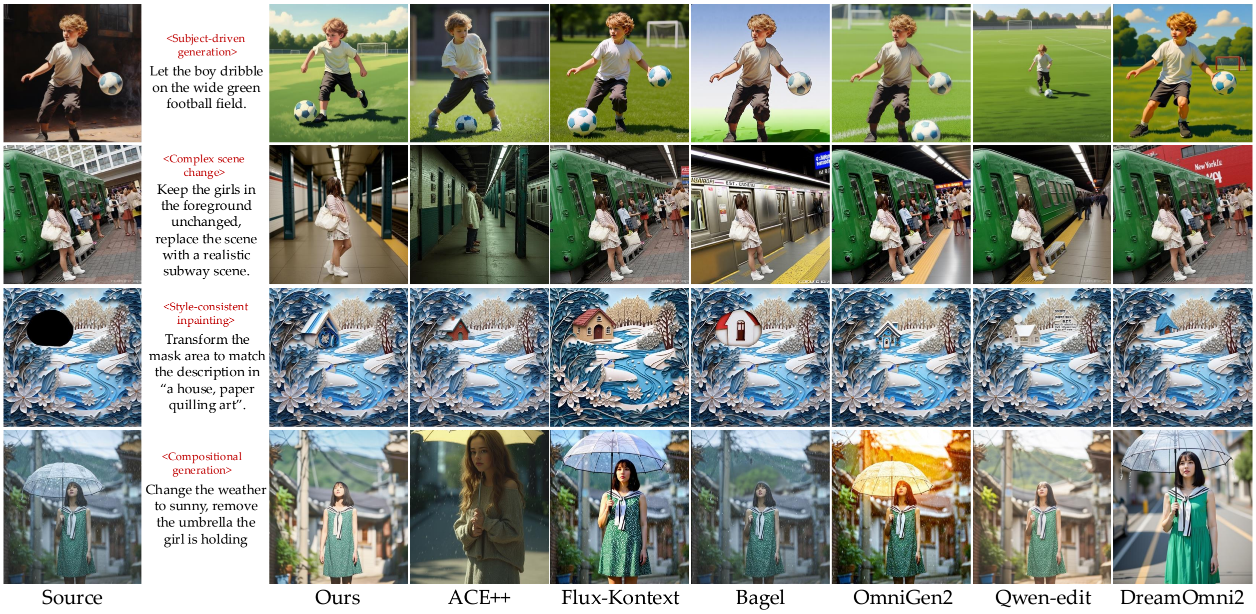

Figure 3. Qualitative comparison on diverse tasks. Our model successfully resolves complex task conflicts where baselines fail.

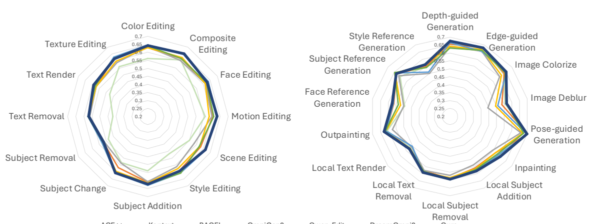

Figure 4. Comprehensive scores on different image editing and generation tasks.

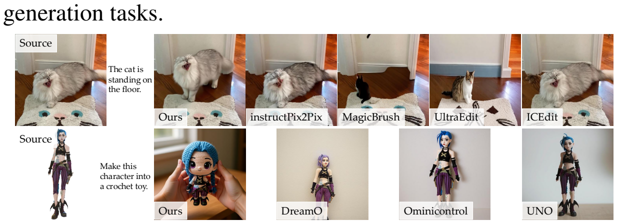

Figure 5. Compare with specialized image editing models and subject-driven generation models. More cases are available in the supplementary metarial.

4.4. Qualitative Comparison

As demonstrated in the preceding qualitative comparison (Fig. 3), our method consistently surpasses SOTA baselines in complex tasks characterized by interfering intents. These unified models typically fail to resolve inherent task conflicts, resulting in critical failures such as "copy-paste" artifacts in subject-driven generation, stylistic dissonance during inpainting, or incomplete execution in compositional editing. Our approach successfully navigates these challenges by utilizing the Predictive Alignment Regularization. This mechanism effectively decouples and routes conflicting sub-tasks (e.g. local semantic edits versus global style preservation) to specialized experts, thereby mitigating the core task interference that plagues unified models.

We further provide comparison against specialized editing and subject-driven models in Fig. 5. As shown in the action editing, the editing models fail to accurately generate realistic poses for action-related edits. Similarly, after modifying material attributes, these subject-driven models struggle to maintain identity consistency. Our model handles both task types effectively. Joint training encourages cross-task benefits: generative diversity from subject data improves action editing, while fidelity constraints from editing data enhance identity preservation. With semantic-aligned routing, the model separates these skills by assigning conflicting objectives to specialized experts, achieving a balance that specialized models cannot.

4.5. Ablation Study

Effectiveness of the MoE architecture. We compare our sparse MoE architecture to a dense baseline of an equivalent activated parameter count. This dense model shows a severe performance drop on ICE-Bench metrics (Tab.4) and slower convergence (Fig. 6 left). This validates that the sparse architecture is fundamentally more effective at mitigating the severe task interference inherent in the unified task space than a computationally-equivalent dense model.

Effect of predictive alignment regularization. We ablate the semantic-alignment loss by removing $\mathcal{L}_{align}$. Without this loss, the MoE gating network performs task-agnostic expert selection, receiving no semantic guidance from our hierarchical tags. As shown in Tab. 4, this variant exhibits substantial degradation across all major metrics. This finding is key: a sparse MoE architecture alone is not sufficient. $\mathcal{L}_{align}$ is what enables semantically guided routing, which is essential for mitigating task interference. No-

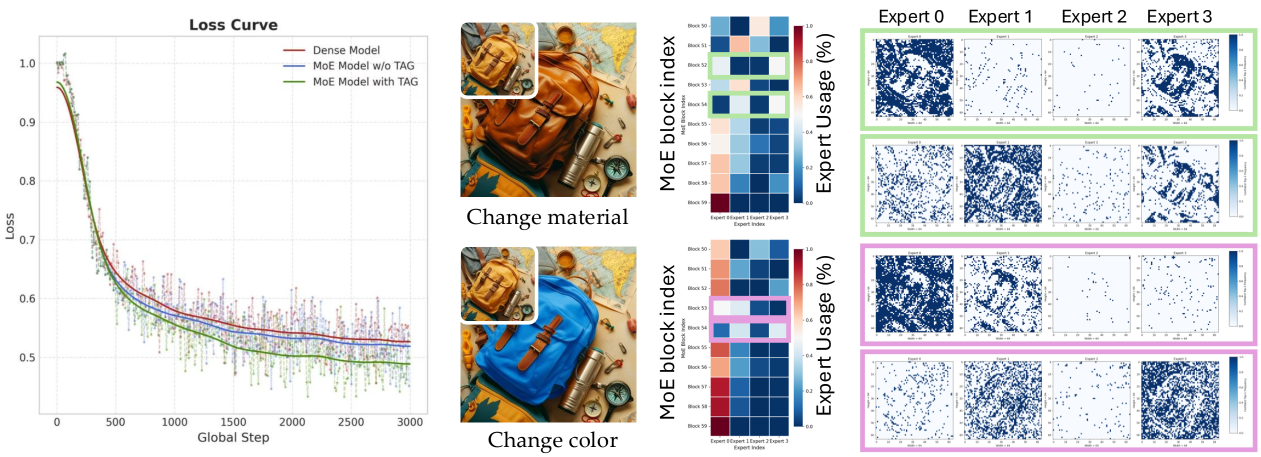

Figure 6. Left: Training loss curves of the dense and MoE architecture. Right: Token strategy in different generation tasks.

tably, the MoE w/o $\mathcal{L}_{align}$ variant still surpasses the dense baseline, benefiting from the larger effective capacity of the sparse MoE structure, which allows exploration of a richer solution space under the same computational budget.

Analysis of expert specialization. To provide direct evidence of our method's success, we visualize the inference-time routing decisions and analyze the internal expert activation patterns. Our analysis is a two-step process. First, we compute an "Expert Utilization Rate" for each MoE layer (shown as the heatmap in the middle of Fig. 6), which represents the percentage of total image tokens routed to each expert. A utilization of 0% (blue) or 100% (red) indicates no specialization. We focus our analysis on layers exhibiting differentiated routing, where utilization is mixed (near white), as this is where functional specialization occurs. Second, for these active layers, we visualize the per-token routing scores for each expert, reshaping them to the image's spatial dimensions. In these token heatmaps, a high score (blue) indicates that the corresponding image tokens are strongly routed to that specific expert. The results reveal a clear, spatially-aware, and task-specific specialization. For Change Material and Change Color, the model activates distinct combinations of experts. Critically, the token heatmaps for these active experts show that computation is spatially concentrated on the backpack's pixels, precisely the region relevant to the edit. The non-relevant background tokens are correctly routed to other experts (or have near-zero activation for these experts). This analysis provides evidence that our model learns a sophisticated, task-specific and spatially-aware specialization, which using distinct expert combinations for different tasks and focusing on semantically relevant image regions, thereby resolving task conflicts by dispatching them to specialized experts.

4.6. User study

We conducted a user study with 65 participants on 50 cases from ICE-Bench [29]. Participants were asked to select the single best result according to three criteria: (1) Reference Alignment (consistency with the source image), (2) Prompt Alignment (faithfulness to the textual instruction), and (3) Overall Preference (overall visual quality). In total, 350 sets were evaluated, and the aggregated results are shown in Fig. 7. The results reveal a clear and consistent preference for our method, which achieved the highest selection rate across all three evaluation criteria.

Method	DINO-ref	Face-ref	Style-ref	CLIP-src	CLIP-cap	vllmqa
Dense	0.7196	0.3544	0.5177	0.851	0.263	0.637
MoE w/o $\mathcal{L}_{align}$	0.7355	0.3779	0.5251	0.863	0.274	0.677
MoE w/ $\mathcal{L}_{align}$	0.7620	0.4642	0.5679	0.879	0.281	0.847

Table 4. Ablation study on dense model and predictive alignment regularization.

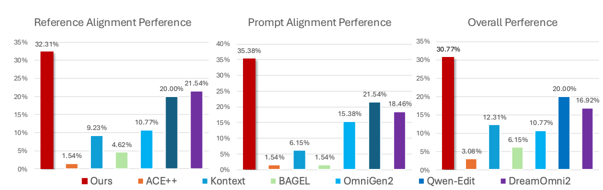

Figure 7. User study on reference alignment, prompt alignment and overall perference.

5. Conclusion

In this paper, we propose TAG-MoE, a task-aware MoE framework for unified image generation and editing. We identify the task-agnostic routing as the core bottleneck for applying MoE to diverse, conflicting tasks. To address this, we introduce a Hierarchical Task Semantic Annotation scheme and Predictive Alignment regularization to effectively injects global task intent into the local routing decisions, forcing the model to develop meaningful expert specialization. Our experiments demonstrate that TAG-MoE significantly mitigates task interference, outperforming dense models and task-agnostic MoE baselines in both quantitative metrics and qualitative fidelity.

Acknowledgements. This work was partly supported by the National Natural Science Foundation of China (NSFC) under Grant 62572458 and the Beijing Major Science and Technology Project under No. Z251100008125009, and National Science and Technology Council, Taiwan under Grant 113-2221-E006-161-MY3.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 4

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 2, 5

[3] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025. 3

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6

[5] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12501–12511, 2025. 2, 5

[6] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021. 5

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2

[8] deepinsight. Insightface. `https://github.com/deepinsight/insightface`, 2021. Accessed: 2025-11-04. 6

[9] DeepSeek-AI. Deepseek-v3 technical report, 2024. 3

[10] 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. 2, 5

[11] Haipeng Fang, Sheng Tang, Juan Cao, Enshuo Zhang, Fan Tang, and Tong-Yee Lee. Attend to not attended: Structure-then-detail token merging for post-training dit acceleration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18083–18092, 2025. 2

[12] Haipeng Fang, Sheng Tang, Zhihao Sun, Ziyao Huang, Juan Cao, Fan Tang, and Yongdong Zhang. Moanimate: Bridging the motion-oriented latent representation gaps in human video animation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2

[13] Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. Scaling diffusion transformers to 16 billion parameters. *arXiv preprint arXiv:2407.11633*, 2024. 2, 3

[14] Tsu-Jui Fu, Yusu Qian, Chen Chen, Wenze Hu, Zhe Gan, and Yinfei Yang. Univg: A generalist diffusion model for unified image generation and editing. *arXiv preprint arXiv:2503.12652*, 2025. 2

[15] Google. Nano banana. Technical report, Google, 2025. 5

[16] Zhen Han, Zeyinzi Jiang, Yulin Pan, Jingfeng Zhang, Chaojie Mao, Chen-Wei Xie, Yu Liu, and Jingren Zhou. Ace: All-round creator and editor following instructions via diffusion transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[17] Runze He, Shaofei Huang, Xuecheng Nie, Tianrui Hui, Luoqi Liu, Jiao Dai, Jizhong Han, Guanbin Li, and Si Liu. Customize your nerf: Adaptive source driven 3d scene editing via local-global iterative training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6966–6975, 2024. 2

[18] Runze He, Kai Ma, Linjiang Huang, Shaofei Huang, Jialin Gao, Xiaoming Wei, Jiao Dai, Jizhong Han, and Si Liu. Freeedit: Mask-free reference-based image editing with multi-modal instruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[19] Runze He, Yiji Cheng, Tiankai Hang, Zhimin Li, Yu Xu, Zijin Yin, Shiyi Zhang, Wenxun Dai, Penghui Du, Ao Ma, et al. Re-align: Structured reasoning-guided alignment for in-context image generation and editing. *arXiv preprint arXiv:2601.05124*, 2026. 2

[20] 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. 2

[21] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 2, 5

[22] Yaowei Li, Lingen Li, Zhaoyang Zhang, Xiaoyu Li, Guangzhi Wang, Hongxiang Li, Xiaodong Cun, Ying Shan, and Yuexian Zou. Blobctrl: A unified and flexible framework for element-level image generation and editing. *arXiv preprint arXiv:2503.13434*, 2025. 2

[23] Yijing Lin, Mengqi Huang, Shuhan Zhuang, and Zhendong Mao. Realgeneral: Unifying visual generation via temporal in-context learning with video models. *arXiv preprint arXiv:2503.10406*, 2025. 2

[24] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 5, 6

[25] Jian Ma, Xujie Zhu, Zihao Pan, Qirong Peng, Xu Guo, Chen Chen, and Haonan Lu. X2edit: Revisiting arbitrary-instruction image editing through self-constructed data and task-aware representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7764–7772, 2026. 2

[26] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487*, 2025. 2, 5

[27] Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. *arXiv preprint arXiv:2504.16915*, 2025. 5

[28] OpenAI. Gpt-4o, 2025. 5

[29] Yulin Pan, Xiangteng He, Chaojie Mao, Zhen Han, Zeyinzi Jiang, Jingfeng Zhang, and Yu Liu. Ice-bench: A unified and comprehensive benchmark for image creating and editing. *arXiv preprint arXiv:2503.14482*, 2025. 5, 6, 8

[30] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3

[31] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 5

[32] Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In *International conference on machine learning*, pages 18332–18346. PMLR, 2022. 3

[33] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2

[34] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024. 5, 6

[35] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024. 6

[36] Yuxin Song, Wenkai Dong, Shizun Wang, Qi Zhang, Song Xue, Tao Yuan, Hu Yang, Haocheng Feng, Hang Zhou, Xinyan Xiao, et al. Query-kontext: An unified multimodal model for image generation and editing. *arXiv preprint arXiv:2509.26641*, 2025. 2

[37] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14940–14950, 2025. 5

[38] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6

[39] Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. Seededit 3.0: Fast and high-quality generative image editing. *arXiv preprint arXiv:2506.05083*, 2025. 5

[40] Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In *The Thirteenth International Conference on Learning Representations*, 2024. 5

[41] 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. 5

[42] 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. 2, 5

[43] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025. 5

[44] Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, et al. Dreamomni2: Multimodal instruction-based editing and generation. *arXiv preprint arXiv:2510.06679*, 2025. 5

[45] 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 Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025. 2

[46] Yu Xu, Fan Tang, Juan Cao, Xiaoyu Kong, Yuxin Zhang, Jintao Li, Oliver Deussen, and Tong-Yee Lee. Headrouter: A training-free image editing framework for mm-dits by adaptively routing attention heads. *ACM Transactions on Graphics*. 2

[47] Yu Xu, Fan Tang, Juan Cao, Yuxin Zhang, Oliver Deussen, Weiming Dong, Jintao Li, and Tong-Yee Lee. B4m: Breaking low-rank adapter for making content-style customization. *ACM Transactions on Graphics*, 44(2):1–17, 2025. 2

[48] Yu Xu, Fan Tang, You Wu, Lin Gao, Oliver Deussen, Hongbin Yan, Jintao Li, Juan Cao, and Tong-Yee Lee. In-context brush: Zero-shot customized subject insertion with context-aware latent space manipulation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–12, 2025. 2

[49] Yu Xu, Yuxin Zhang, Juan Cao, Lin Gao, Chunyu Wang, Oliver Deussen, Tong-Yee Lee, and Fan Tang. Beyond pix-

els: Visual metaphor transfer via schema-driven agentic reasoning. *arXiv preprint arXiv:2602.01335*, 2026. 2

[50] Zijin Yin, Tiankai Hang, Yiji Cheng, Shiyi Zhang, Runze He, Yu Xu, Chunyu Wang, Bing Li, Zheng Chang, Kongming Liang, et al. Generative visual chain-of-thought for image editing. *arXiv preprint arXiv:2603.01893*, 2026. 2

[51] 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*, 36:31428–31449, 2023. 5

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 5

[53] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (ToG)*, 42(6):1–14, 2023. 2

[54] 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. 3

[55] 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*, 37:3058–3093, 2024. 5

[56] Youwei Zheng, Yuxi Ren, Xin Xia, Xuefeng Xiao, and Xiaohua Xie. Dense2moe: Restructuring diffusion transformer to moe for efficient text-to-image generation. *arXiv preprint arXiv:2510.09094*, 2025. 2, 3

[57] Tianrui Zhu, Shiyi Zhang, Jiawei Shao, and Yansong Tang. Kv-edit: Training-free image editing for precise background preservation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16607–16617, 2025. 2