INSTRUCTMOLE: INSTRUCTION-GUIDED MIXTURE OF LOW-RANK EXPERTS FOR MULTI-CONDITIONAL IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Parameter-Efficient Fine-Tuning of Diffusion Transformers (DiTs) for diverse, multi-conditional tasks often suffers from task interference when using monolithic adapters like LoRA. The Mixture of Low-rank Experts (MoLE) architecture offers a modular solution, but its potential is usually limited by routing policies that operate at a token level. Such local routing can conflict with the global nature of user instructions, leading to artifacts like spatial fragmentation and semantic drift in complex image generation tasks. To address these limitations, we introduce InstructMoLE, a novel framework that employs an Instruction-Guided Mixture of Low-Rank Experts. Instead of per-token routing, InstructMoLE utilizes a global routing signal, Instruction-Guided Routing (IGR), derived from the user's comprehensive instruction. This ensures that a single, coherently chosen expert council is applied uniformly across all input tokens, preserving the global semantics and structural integrity of the generation process. To complement this, we introduce an output-space orthogonality loss, which promotes expert functional diversity and mitigates representational collapse. Extensive experiments demonstrate that InstructMoLE significantly outperforms existing LoRA adapters and MoLE variants across challenging multi-conditional generation benchmarks. Our work presents a robust and generalizable framework for instruction-driven fine-tuning of generative models, enabling superior compositional control and fidelity to user intent.

1 INTRODUCTION

The advent of powerful, open-source Diffusion Transformers (DiTs) (Peebles & Xie, 2023; Esser et al., 2024; Labs, 2024) has unlocked unprecedented capabilities in generative AI, fueling a demand for highly specialized and compositional functionalities, from multi-subject composition to personalized content creation (Labs et al., 2025; Wu et al., 2025a;b; Liu et al., 2025a). Parameter-Efficient Fine-Tuning (PEFT), particularly Low-Rank Adaptation (LoRA) (Hu et al., 2022), has become the de facto standard for such customization. However, LoRA's monolithic update structure conflicts with the demands of multi-task fine-tuning, leading to catastrophic forgetting as different task objectives interfere (Biderman et al., 2024; Han et al., 2024).

The Mixture-of-Experts (MoE) architecture offers a structured solution to this interference problem. As theoretically grounded by (Li et al., 2025), MoE mitigates catastrophic forgetting by diversifying its experts to specialize in different tasks, which in turn helps to establish, or at least preserve, task-specific expert circuits and balance the loads across them. In the domain of language and multi-modal models, MoE has been extensively studied, leading to established, often task-aware, routing strategies (Wu et al.; Liu et al., 2025c; Li et al., 2024; Gou et al., 2023; Chen et al.; Wu et al., 2025c; Sun et al., 2025; Fei et al., 2024).

In contrast, its application to diffusion transformers for multi-conditional image generation (e.g. image editing, multi-subject driven generation) remains comparatively underexplored. Initial works like ICEdit (Zhang et al., 2025b) have adopted conventional token-level routing mechanisms without specializing them for this new problem context. This reliance on an ill-suited, token-level routing mechanism reveals a misalignment with the nature of image generation instructions. A user's in-

struction for compositional generation, for instance, to create a scene with "*A baby crawling on the grass, a white horse grazing nearby, and a football helmet*" (in Figure 2), establishes a complex set of global semantic relationships. In contrast, token-level routing delegates expert selection to each local image patch independently. This flawed, uncoordinated decision-making process is the primary cause of critical failures, often resulting in subjects that are incoherently rendered, spatially fragmented, or have their specified relationships and attributes ignored entirely.

To rectify this misalignment, we propose InstructMoLE, a MoLE framework for multi-conditional image generation built upon a novel, globally consistent routing policy. The core of our framework is Instruction-Guided Routing (IGR), which conditions expert selection entirely on the global semantics of the user's textual instruction.This mechanism ensures that for any given layer of the model, a single, unified "expert council" is chosen based on the instruction and broadcast to all spatial locations within that layer. This enforces processing consistency at each stage of generation, while critically allowing the model to recruit different sets of specialized experts across different layers, tailoring the computation to the varying levels of abstraction. However, the effectiveness of this globally-applied council hinges on the functional diversity of its constituent experts. To ensure this diversity and prevent representational collapse, we complement IGR with a novel output-space orthogonality loss. This regularizer encourages the learned experts to occupy distinct functional roles, thereby maximizing the compositional power of the selected council. Our main contributions are:

- We identify and address a critical challenge in applying MoLE to instruction-based editing and generation: the inherent mismatch between local, token-level routing policies and the global, semantic scope of user instructions. To address this, we propose InstructMoLE, a framework centered on Instruction-Guided Routing (IGR), which aligns expert selection with the holistic intent of the instruction.
- We introduce a novel output-space orthogonality regularizer to explicitly promote functional diversity among experts. This technique complements standard load-balancing losses by directly penalizing representational redundancy in the expert outputs, thereby mitigating expert collapse and improving the model's compositional control.
- We provide extensive empirical validation across multiple challenging benchmarks for multi-conditional image generation. Our results demonstrate that InstructMoLE achieves state-of-the-art performance, particularly in tasks demanding high compositional fidelity and adherence to complex spatial relationships. This directly validates the superiority of our global, instruction-guided approach over traditional token-level routing policies. These findings validate that instruction-guided routing is a more effective and principled approach for multi-conditional generation tasks.

2 Related Work

2.1 Conditional Generation with Diffusion Transformers.

The advent of Diffusion Transformers (DiTs) has marked a new era for generative modeling, demonstrating remarkable scalability and performance (Peebles & Xie, 2023; Esser et al., 2024). A significant line of research has focused on enhancing their controllability for complex, instruction-driven tasks. Methods like DreamO (Mou et al., 2025) and In-Context Edit (Zhang et al., 2025b) have enabled sophisticated image customization and editing based on diverse user inputs. Others have introduced lightweight modules for flexible conditioning (Zhang et al., 2025a; Jiang et al., 2025) or tackled data bottlenecks in multi-subject scenarios (Wu et al., 2025d). While these works significantly advance the state of conditional generation, they typically rely on a monolithic adaptation strategy, such as a single LoRA (Hu et al., 2022). This approach struggles when tasked with mastering multiple, potentially conflicting skills simultaneously, creating a clear need for more dynamic, modular adaptation methods.

2.2 Mixture-of-Experts for Parameter-Efficient Fine-Tuning.

The Mixture-of-Experts (MoE) architecture, particularly in its low-rank form (MoLE), offers a powerful solution to the limitations of monolithic fine-tuning. By allocating specialized LoRA experts

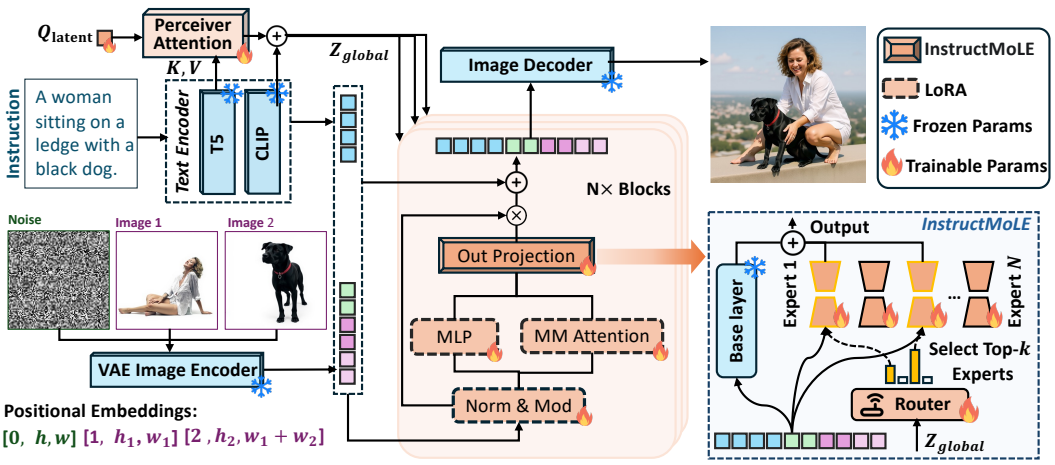


Figure 1: Illustration of the InstructMoLE framework. A global signal, $\mathbf{Z}_{\text{global}}$, is distilled from the user's instruction to guide a Router. The Router selects a single, consistent set of LoRA experts, which is then applied to all input tokens.

for different functions, MoLE can mitigate task interference and enhance model capacity efficiently. This paradigm has been extensively and successfully explored in language and multi-modal models for multi-task learning and instruction tuning (Dou et al., 2024; Gou et al., 2023; Chen et al.).Its application to vision, however, is more nascent. While initial works have validated the potential of MoLE for tasks like controllable visual effect generation (Mao et al., 2025), a well-designed routing policy that determines which experts to activate for a given input is crucial for fully harnessing the power of this architecture.

2.3 MoE Design: Routing Policies and Expert Diversity.

The design of an effective MoE system centers on two core challenges: the routing policy and the promotion of expert diversity. The predominant routing approach, inherited from language models, is token-level routing, where each token independently selects experts (Shazeer et al., 2017; Zhang et al., 2025b). While advanced variants like Expert Choice routing improve dynamics (Sun et al.), they inherit the same fundamental limitation: local, per-token decision-making. This is misaligned with the holistic intent of user instructions in image generation. A second, orthogonal challenge is the common issue of "expert collapse", where experts become functionally redundant. Standard load-balancing only encourages uniform utilization, not functional diversity. Our work addresses these dual challenges head-on. We introduce Instruction-Guided Routing (IGR), a global policy that aligns expert selection with the user's instruction, and complement it with an output-space orthogonality loss that directly enforces functional diversity among experts.

3 Methodology

We present InstructMoLE, a Mixture of Low-rank Experts (MoLE) framework for multi-conditional image generation, illustrated in Figure 1. The framework champions a singular, top-down routing principle: a unified Instruction-Guided Routing (IGR) policy is applied consistently across all model layers. This is achieved through a Perceiver-style signal distillation for robust global guidance and an output-space orthogonality loss to maintain functional diversity among experts.

3.1 InstructMoLE Architecture

Standard LoRA augments a frozen linear layer weight $\mathbf{W}_0 \in \mathbb{R}^{D_{\text{in}} \times D_{\text{out}}}$ with a single, static low-rank update. We replace this static update with a dynamic, sparse mixture of N expert LoRA modules $\{\mathbf{E}_i\}_{i=1}^{N}$. Each expert i consists of a down-projection matrix $\mathbf{W}_{\text{A}}^i \in \mathbb{R}^{D_{\text{in}} \times r}$ and an up-projection matrix $\mathbf{W}_{\text{B}}^i \in \mathbb{R}^{r \times D_{\text{out}}}$, where $r \ll D_{\text{in}}, D_{\text{out}}$ is the rank. The function for expert i is defined as

$\mathbf{E}_i(\mathbf{X}) = \mathbf{X}\mathbf{W}_{\text{A}}^i\mathbf{W}_{\text{B}}^i$. The output of a MoLE layer is therefore:

$$\text{MoLE}(\mathbf{X}) = \mathbf{X}\mathbf{W}_0 + \sum_{i=1}^{N} g_i(\cdot)\mathbf{E}_i(\mathbf{X}) \tag{1}$$

where the gating weights $g_i(\cdot)$ are determined by a routing network; $\mathbf{X} \in \mathbb{R}^{B\times L\times D_{\text{in}}}$ is the input tensor with batch size B and sequence length L.

Token-level vs. Instance-level Routing. Conventional MoE models employ *token-level* routing, where each of the $B \times L$ tokens independently selects experts from its local hidden state (Fedus et al., 2022; Shazeer et al., 2017; Yuan et al.; Sun et al.; Zhang et al., 2025b), resulting in a dense logits tensor of shape $\mathbb{R}^{B\times L\times N}$. The uncoordinated, per-token decisions create a tension with the inherently global nature of many instructions, leading to critical failures: (i) spatial fragmentation, yielding region-wise style and hue discontinuities; (ii) amplification of high-frequency noise through routing jitter; and (iii) weakened adherence to the user instruction.

To counteract these failure modes, we introduce an *instance-level* policy that enforces a globally consistent expert assignment at each layer. Instead of fragmented per-token decisions, our router is guided by the global semantics of the user instruction to compute a *unified expert council* for the entire instance. This council, represented by a compact logits tensor of shape $\mathbb{R}^{B\times 1\times N}$, is then broadcast identically to all L tokens within that layer. We term this mechanism Instruction-Guided Routing. By ensuring that a single, unified expert council is applied to all input tokens within a given layer, IGR directly prevents fragmentation and routing jitter, thereby preserving the semantic and structural integrity of the image.

3.1.1 Instruction-Guided Routing (IGR) Policy

Global Routing Signal ($\mathbf{Z}_{\text{global}}$). The IGR policy is conditioned on a global signal, $\mathbf{Z}_{\text{global}}$, which is distilled from the editing instruction, $\mathbf{I_c}$. To construct a signal that is both semantically robust and compositionally aware, we fuse two distinct text representations. Given an $\mathbf{I_c}$, a T5 encoder produces token-level features $\mathbf{H}_{\text{inst}} \in \mathbb{R}^{B\times L_{\text{inst}}\times D_{\text{inst}}}$, where B is the batch size, L_{inst} is the instruction length, and D_{inst} is the T5 embedding dimension. Concurrently, a CLIP encoder provides a pooled embedding $\text{CLIP}(\mathbf{I_c}) \in \mathbb{R}^{B\times D}$. While the pooled embedding offers a strong semantic anchor, it can overlook crucial compositional nuance. For example, in the instruction "Change the woman's dress to red and the man's shirt to blue", a pooled embedding might average "red" and "blue", losing the critical association between the colors and their respective subjects. Conversely, token-level features, if naively averaged, risk diluting these key semantics.

We therefore employ a **Perceiver-style attentional bottleneck** (Jaegle et al., 2021) to distill the rich, token-level information from $\mathbf{H}_{\text{inst}}$ into a compact summary before fusing it with the CLIP embedding. Concretely, we first project the T5 tokens into the D-dimensional CLIP space using a linear layer $\mathbf{W}_{\text{in}} \in \mathbb{R}^{D_{\text{inst}}\times D}$. A single, learnable latent query, $\mathbf{Q}_{\text{latent}} \in \mathbb{R}^{1\times 1\times D}$, is then used to iteratively query the projected instruction tokens via $S = 2$ layers of Perceiver-style attention:

$$\begin{aligned} \tilde{\mathbf{X}} \;&=\; \mathbf{H}_{\text{inst}}\mathbf{W}_{\text{in}}, \quad \mathbf{L}^{(0)} \;=\; \text{tile}_B(\mathbf{Q}_{\text{latent}}), \\ \mathbf{L}^{(s)} \;&=\; \mathbf{L}^{(s-1)} + \text{PerceiverAttn}(Q = \mathbf{L}^{(s-1)}, K = \tilde{\mathbf{X}}, V = \tilde{\mathbf{X}}), \end{aligned} \tag{2}$$

where $\text{tile}_B(\cdot)$ repeats the latent query B times. The final latent, $\mathbf{L}^{(S)}$, now serves as a distilled summary of the compositional details present in the T5 features. This summary is projected with $\mathbf{W}_{\text{out}} \in \mathbb{R}^{D\times D}$, normalized, and then additively fused with the CLIP embedding:

$$\mathbf{Z}_{\text{global}} \;=\; \underbrace{\text{LayerNorm}(\mathbf{L}^{(S)}\mathbf{W}_{\text{out}})}_{\text{distilled compositional summary}} + \underbrace{\text{CLIP}(\mathbf{I_c})}_{\text{holistic semantics}}. \tag{3}$$

This design allows the distilled term to contribute fine-grained specificity from the instruction, while the CLIP embedding provides a robust, holistic semantic anchor, together forming a balanced and powerful global routing signal.

IGR Gating Mechanism. The IGR gating mechanism is instantiated independently at each MoLE layer. We formalize the computation by first describing the process for a single instance

$b \in \{1, \ldots, B\}$ within a batch, which corresponds to the input tensor slice $\mathbf{X}_b \in \mathbb{R}^{L \times D_{\text{in}}}$. The per-instance outputs, $\mathbf{Y}_b$, are subsequently stacked to form the full batch output $\mathbf{Y}$.

At a given l-th layer, the gating network $\mathcal{G}_l : \mathbb{R}^D \to \mathbb{R}^N$ generates a logit vector over the N experts from a shared global instruction signal, $\mathbf{Z}_{\text{global},b}$. The Top-$k$ experts are then selected:

$$(\mathcal{I}_b, \mathbf{w}_b) = \text{Top-}k(\text{Softmax}(\mathcal{G}_l(\mathbf{Z}_{\text{global},b})), k). \tag{4}$$

Here, $\mathcal{I}_b \subset \{1, \ldots, N\}$ is the set of indices for the Top-k experts, and $\mathbf{w}_b \in \mathbb{R}^k$ contains their associated, unnormalized Softmax probabilities. $\mathcal{I}_b$ and $\mathbf{w}_b$ is then broadcast across all L token positions in the sequence. The final output for the l-th layer, $\mathbf{Y}_b$, is computed by summing the output of a shared linear layer with the weighted sum of the selected expert outputs:

$$\mathbf{Y}_b = \mathbf{X}_b \mathbf{W}_0 + \sum_{j=1}^{k} w_{b,j} \cdot \mathbf{E}_{\mathcal{I}_b[j]}(\mathbf{X}_b), \tag{5}$$

where $w_{b,j}$ is the weight for the j-th selected expert, whose global index is $\mathcal{I}_b[j]$), and $\mathbf{E}_{\mathcal{I}_b[j]}$ denotes the corresponding low-rank expert.

3.2 Training Objectives

Our training objective is designed to achieve two goals simultaneously: high-fidelity image generation and the effective, diverse utilization of the expert pool. This is realized through a composite loss function, where the λ terms are hyper-parameters:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{flow}} + \lambda_{\text{aux}} \mathcal{L}_{\text{aux}} + \lambda_{\text{ortho}} \mathcal{L}_{\text{ortho}}. \tag{6}$$

Generation Accuracy ($\mathcal{L}_{\text{flow}}$). The primary objective, $\mathcal{L}_{\text{flow}}$, is the standard flow-matching or diffusion loss of the MM-DiT backbone. This ensures the model accurately synthesizes images that are well aligned with the user's instruction.

Expert Load Balancing ($\mathcal{L}_{\text{aux}}$). Following the standard practice for training MoEs (Shazeer et al., 2017), we employ an auxiliary load-balancing loss, $\mathcal{L}_{\text{aux}}$, to prevent the model from consistently favoring only a few experts. This loss is a function of two quantities calculated over a batch: f_i, the fraction of tokens routed to expert i, and p_i, the mean routing probability for that expert:

$$\mathcal{L}_{\text{aux}} \;=\; N \sum_{i=1}^{N} f_i \cdot p_i \,, \tag{7}$$

where N is the total number of experts.

Output-space Orthogonality Loss for Functional Diversity ($\mathcal{L}_{\text{ortho}}$). A critical failure mode in Mixture-of-Experts (MoE) models is *expert collapse* (Fedus et al., 2022), where distinct experts converge to functionally redundant solutions, thereby nullifying the benefits of the mixture. The standard mitigation, an auxiliary load-balancing loss, only encourages that experts are utilized with similar frequency but provides no explicit mechanism to ensure their functional diversity.

To address this limitation directly, we introduce an **orthogonality loss** that penalizes the functional similarity between pairs of experts. Our approach operates on the raw, pre-gating outputs. For a given input batch $\mathbf{X}$, we first compute the output of every expert, yielding a set of tensors $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_N\}$, where each $\mathbf{Y}_i = \mathbf{E}_i(\mathbf{X}) \in \mathbb{R}^{B \times L \times D_{\text{out}}}$. To measure the functional similarity between any two experts i and j, we flatten their respective output tensors into high-dimensional vectors, $\mathbf{v}_i = \text{vec}(\mathbf{Y}_i)$ and $\mathbf{v}_j = \text{vec}(\mathbf{Y}_j)$.

The orthogonality loss, $\mathcal{L}_{\text{ortho}}$, is then defined as the mean of the squared cosine similarities over all unique pairs of these expert output vectors:

$$\mathcal{L}_{\text{ortho}} = \frac{1}{N(N-1)} \sum_{i \neq j} \left(\frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\|_2 \|\mathbf{v}_j\|_2} \right)^2. \tag{8}$$

Minimizing this objective incentivizes the vectors $\{\mathbf{v}_i\}_{i=1}^N$ to become mutually orthogonal. This, in turn, forces the expert functions $\{\mathbf{E}_i\}_{i=1}^N$ to learn distinct and complementary representations,

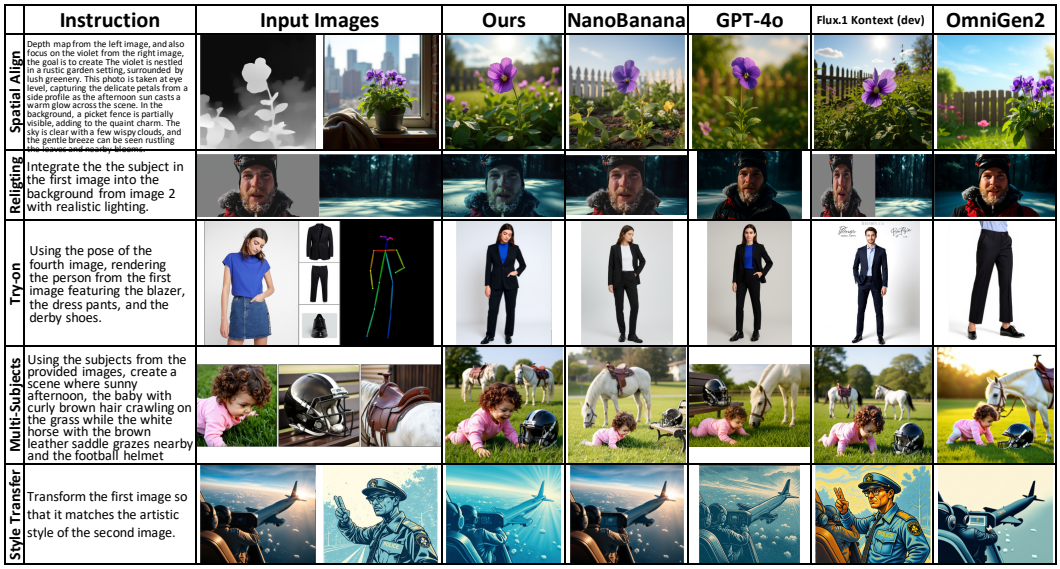


Figure 2: Qualitative comparison with state-of-the-art models.

effectively preventing expert collapse. In practice, this loss is computed efficiently by calculating the squared off-diagonal elements of the Gram matrix formed by the L2-normalized vectors $\{\mathbf{v}_i/\|\mathbf{v}_i\|_2\}_{i=1}^{N}$. This is crucial for unlocking the full specialization capacity of our MoLE framework.

4 Experiments

4.1 Experimental Setup

Training Data. Our model's versatile editing capabilities are a direct result of its comprehensive training mixture. We curate a large-scale dataset by combining publicly available sources with a vast corpus of synthesized data, designed to expose the model to a diverse spectrum of conditional inputs. As illustrated in Figure 6, this includes reference-based tasks (e.g., face swapping, style transfer, re-lighting), multi-subject compositional generation, single-image editing, and spatially controlled generation from both dense (depth, Canny maps) and pose skeleton signals. This diverse training regimen enables a single, unified model to handle a wide array of editing modalities.

Evaluation Benchmarks. We evaluate all models on a suite of benchmarks, each targeting a distinct capability: OmniContext (Chen & Gupta, 2025) for in-context generation, XVerseBench (Chen et al., 2025) for multiple subjects-driven generation, GEdit-EN-full (Liu et al., 2025b) for single-image editing, and MultiGen-20M (Qin et al., 2023) and COCO Pose 2017 (Lin et al., 2014) for spatially controlled generation.

In the following tables, the best results are in **bold** and the second best are underlined. (↓: Lower is better; ↑: Higher is better).

4.2 Main Results

Implementation Details. We fine-tune our model from the Flux.1 Kontext (dev) backbone for 100K steps on 64 NVIDIA H100 GPUs. For the MoLE layers, we use $N = 8$ total experts with $k = 4$ activated per token and an expert rank of $r = 32$. The model is optimized using AdamW with a constant learning rate of 1×10^{-4} and a global batch size of 256 (4 per device).

Qualitative Comparison. Figure 2 qualitatively compares our model against state-of-the-art methods on several complex editing tasks, visually confirming its superiority. In tasks requiring

Table 1: Quantitative comparison on the OmniContext benchmark, evaluating both prompt following and identity similarity (ID-Sim), with corresponding qualitative results shown in Figure 3.

Method	Prompt Following	ID-Sim					
		Avg.	multi character	multi character object	scene character	scene character object	single character
UNO	5.63	16.56	20.65	15.16	16.10	8.49	22.41
DreamO-v1.1	6.10	15.47	19.30	9.46	12.41	8.42	27.77
OmniGen2	**7.58**	23.81	27.52	18.68	18.20	12.61	42.02
Flux.1 Kontext (dev)	6.24	<u>36.29</u>	**50.08**	<u>31.66</u>	<u>30.10</u>	<u>14.97</u>	<u>54.65</u>
InstructMoLE	<u>6.75</u>	**38.85**	<u>45.12</u>	**34.03**	**34.59**	**20.04**	**60.47**

Figure 3: Qualitative comparison of in-context generation on OmniContext benchmark.

precise spatial control like *Spatial Align* and *Try-on*, competing models often fail to respect geometric constraints, yielding misaligned objects or incorrect poses. In contrast, our model robustly adheres to both dense (depth) and sparse (pose) guidance. Our method also demonstrates a stronger grasp of compositional semantics and identity preservation, particularly in the *Multi-Subjects* task where it correctly renders all subjects and their relationships. This advantage is most pronounced in the *Swap Face* task, where InstructMoLE is the only model to produce a coherent result while others fail. This ability to disentangle and execute the spatial, semantic, and identity components of a complex instruction highlights the effectiveness of our approach.

In-Context Generation. As shown in Table 1, InstructMoLE achieves a new state-of-the-art on the OmniContext benchmark, attaining the highest average ID-Sim score. The benchmark evaluates instruction adherence (assessed by GPT-4.1) and identity preservation (ID-Sim) (Deng et al., 2019). While OmniGen2 scores highest in *Prompt Following*, it suffers from poor identity preservation, scoring 39% lower than InstructMoLE (Ours) on the ID-Sim average. InstructMoLE demonstrates a superior balance, outperforming the strong Flux.1 Kontext baseline on both metrics. Although Flux.1 Kontext scores highest on the *multi character* sub-metric, this is due to rendering artifacts where subjects are naively concatenated rather than composed into a coherent scene, as shown in Figure 3 and Figure 9.

Table 2: Quantitative comparison of multi-subject driven generation. Corresponding qualitative results are presented in Figure 4.

Method	DPG	ID-Sim	IP-Sim	Avg.
UNO	74.35	31.96	53.07	53.13
DreamO-v1.1	<u>89.92</u>	<u>58.86</u>	59.93	<u>69.57</u>
OmniGen2	**91.91**	35.95	<u>60.93</u>	62.93
Flux.1 Kontext (dev)	89.49	53.12	59.78	67.46
InstructMoLE	89.57	**60.84**	**62.81**	**71.07**

Multiple Subjects-driven Generation. We evaluate the model's ability to compose multiple subjects using XVerseBench (Chen et al., 2025), a benchmark comprising 20 human identities, 74 unique objects, and 45 different animal species. As detailed in Table 2, we assess performance using three key metrics: instruction adherence (DPG score (Hu et al., 2024)), human identity preservation (Face ID similarity (Deng et al., 2019)), and object fidelity (DINOv2 similarity (Oquab et al., 2024)).

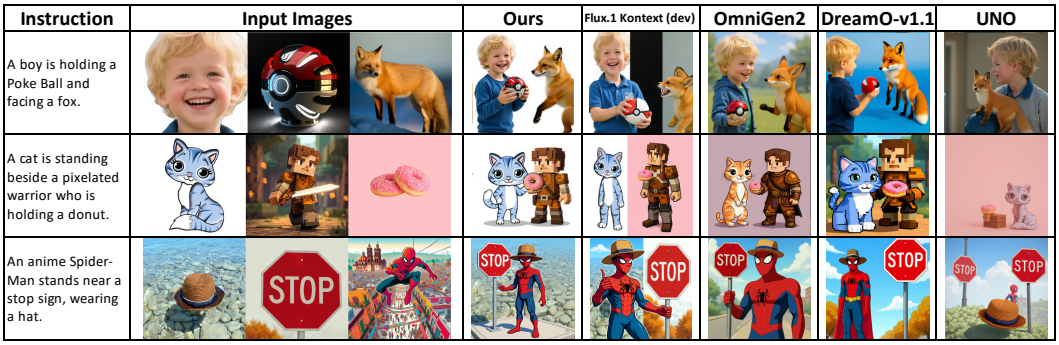


Figure 4: Qualitative comparison of multi-subject driven generation on XVerse benchmark.

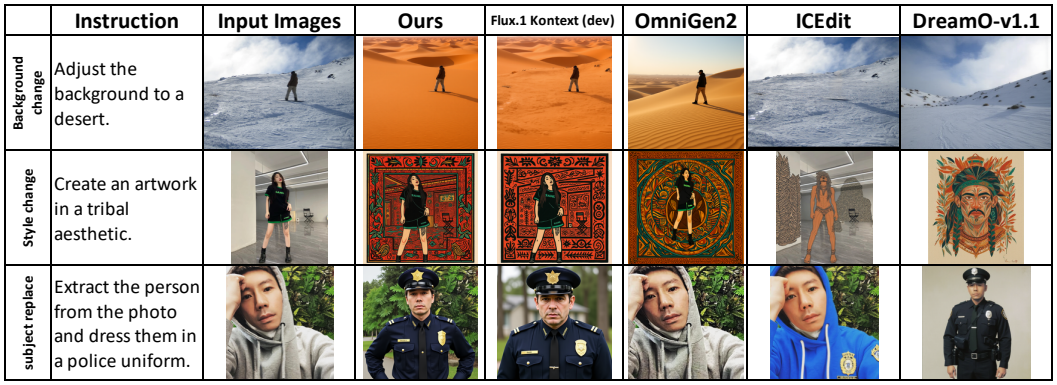


Figure 5: Qualitative comparison of single-image editing on GEdit-EN-full benchmark.

InstructMoLE achieves a new state-of-the-art on the benchmark. It outperforms the strong Flux.1 Kontext baseline on the average score by a significant margin of over 3.6. While its instruction adherence (DPG) is on par with other leading methods, our model establishes a decisive advantage in subject fidelity, securing the top scores in both identity preservation (ID-Sim) and object similarity (IP-Sim). More qualitative comparison results are presented in Figure 10.

Single-image Editing. As shown in Table 3, on the GEdit-EN-full benchmark, which evaluates performance across 11 distinct real-world editing categories by GPT-4.1, InstructMoLE again demonstrates its superior versatility by achieving the highest average score. Our model surpasses the strong Flux.1 Kontext baseline and exhibits a decisive advantage over other leading methods like OmniGen2, with an overall performance improvement of more than 13% over the latter. This strength is particularly evident in its ability to handle both global and local manipulations with high fidelity. It establishes a new state-of-the-art in fundamental tasks such as *Color Alter*, *Material Alter*, and *Replace*, while remaining highly competitive across nearly all other categories.

Table 3: Quantitative comparison of single-image editing. The benchmark comprises 11 fine-grained editing categories reflecting practical user requests, with performance for each assessed by GPT-4.1. The corresponding qualitative results shown in Figure 5.

Model	BG Change	Color Alt.	Mat. Alt.	Motion	Portrait	Style	Add	Remove	Replace	Text	Tone	Avg
DreamO-v1.1	3.06	1.66	2.35	3.76	3.24	3.34	2.16	0.55	3.02	1.51	2.06	2.43
ICEdit	2.73	6.00	4.41	1.74	2.14	5.19	4.41	1.53	4.22	1.58	4.58	3.50
OmniGen2	6.99	5.10	5.11	**3.93**	**4.59**	**6.88**	6.17	4.68	6.45	4.04	6.05	5.45
Flux.1 Kontext (dev)	6.99	7.17	5.60	3.13	4.29	6.70	6.90	**6.92**	6.27	**5.56**	7.14	6.06
InstructMoLE	**7.03**	**7.46**	**5.81**	3.29	4.20	6.79	**7.03**	6.87	**6.53**	5.43	**7.42**	**6.17**

4.3 Ablation Studies

Implementation Details. We train all models in our ablation studies under a unified and consistent experimental protocol to ensure a fair comparison, using 8 NVIDIA H100 GPUs for 20K steps.

Evaluation Metrics. We conduct a comprehensive evaluation across three key domains. For *multiple subjects-driven generation* and *single-image editing*, we report scores on the XVerseBench (Chen et al., 2025), GEdit-EN-full (Zhang et al., 2024). For *spatial alignment*, we evaluate on 500 samples for each modality, randomly sampled from the evaluation splits of MultiGen-20M (Qin et al., 2023) for depth and Canny edge control, and COCO Pose 2017 (Lin et al., 2014) for human pose control. The corresponding metrics are Root Mean Squared Error (RMSE) for depth, F1 score for Canny edges, and an F1 score based on Object Keypoint Similarity (OKS) for pose.

Table 4: Ablation study on routing policies. "Standard" refers to the classic token-level Top-k routing policy (Fei et al., 2024). "Standard+IGR" applies the Standard policy to the early dual-stream blocks and IGR to the later single-stream blocks. "IGR+Standard" applies the reverse configuration.

Model	Routing Policy	Multi-Subject (↑)	Single-Subject (↑)	Depth MSE (↓)	Canny F1 (↑)	Pose F1 (↑)
Flux.1 Kontext (dev)	-	63.87	6.06	63.44	13.72%	0.95%
LoRA ($r = 256$)	-	64.68	6.06	35.92	40.43%	31.94%
MoLE ($r = 32$, $N = 8$, $k = 4$)	Standard	64.56	6.07	38.64	28.06%	19.50%
	Expert Choice (EC)	**65.81**	6.07	34.00	38.55%	31.47%
	Expert Race (ER)	55.83	6.11	102.67	16.73%	0.00%
	IGR+Standard	65.17	**6.15**	33.43	38.33%	33.77%
	Standard+IGR	64.86	6.07	36.32	33.33%	9.57%
	IGR (ours)	65.68	**6.15**	**33.34**	**41.51%**	**40.97%**

Table 5: Ablation study of the IGR signal and the orthogonality loss.

IGR Signal	Orthogonality Loss	Multi-Subject (↑)	Single-Subject (↑)	Depth MSE (↓)	Canny F1 (↑)	Pose F1 (↑)
CLIP($\mathbf{I_c}$)	w/o	64.54	6.12	35.67	38.12%	36.74%
$\mathbf{Z}_{\text{global}}$	w/o	65.65	5.97	33.74	38.92%	37.77%
$\mathbf{Z}_{\text{global}}$	w/	**65.68**	**6.15**	**33.34**	**41.51%**	**40.97%**

Routing Policy Analysis. We compare our Instruction-Guided Routing (IGR) against state-of-the-art token-level policies in Table 4, using a high-rank LoRA as a strong baseline under a fair training parameter budget. The results reveal the inherent instability of token-level routing for this domain; the standard Top-k policy degrades performance, and this failure is exacerbated by the flexible Expert Race (ER) (Sun et al.), which collapses entirely on the Pose F1-task. While the load-balanced Expert Choice (EC) (Yuan et al.) proves to be a more stable token-level alternative, it is consistently outperformed by our IGR policy. IGR's superiority demonstrates that for multi-conditional generation, enforcing a coherent, instance-level signal is more critical than the adaptive, token-wise compute allocation offered by even the strongest token-level methods.

Analysis of IGR Components. We conduct a fine-grained ablation on the core components of our IGR policy in Table 5. The results provide two clear validations. First, replacing the raw CLIP embedding with our fused signal, $\mathbf{Z}_{\text{global}}$, significantly enhances performance across nearly all metrics, confirming the necessity of distilling compositional details from token-level features for precise control. Second, building upon this superior signal, the addition of our orthogonality loss yields another substantial boost, particularly in spatial alignment, increasing the Canny F1 score by over 2.5 and the Pose F1 score by over 3 percentage points. This confirms that both our signal distillation and our orthogonality regularizer are critical and complementary components for achieving state-of-the-art performance.

5 Conclusion

In this work, we addressed multi-task interference in parameter-efficient diffusion models by resolving the conflict between local token-level routing and the global intent of user instructions. We introduced InstructMoLE, a framework built upon Instruction-Guided Routing (IGR). This policy enforces a globally consistent expert selection within each layer based on the user's instruction, a mechanism complemented by an output-space orthogonality loss that promotes expert diversity. Our extensive experiments show that InstructMoLE significantly outperforms LoRA and other MoLE variants on challenging multi-conditional benchmarks. Ultimately, our work validates that a global, instruction-aware routing policy is a more robust and effective paradigm for complex generative tasks, laying the groundwork for models with more faithful compositional control.

REFERENCES

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. Lora learns less and forgets less, 2024. URL https://arxiv.org/abs/2405.09673.

Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen, Xu Wang, Kang Du, and Xinglong Wu. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. *arXiv preprint arXiv:2506.21416*, 2025.

Hao Chen and Anjali Gupta. OmniContext: A testbed for context-aware generative editing. *arXiv preprint arXiv:2501.05678*, 2025.

Zeren Chen, Ziqin Wang, Zhen Wang, Huayang Liu, Zhenfei Yin, Si Liu, Lu Sheng, Wanli Ouyang, and Jing Shao. Octavius: Mitigating task interference in mllms via lora-moe. In *The Twelfth International Conference on Learning Representations*.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, et al. Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1932–1945, 2024.

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 *Forty-first international conference on machine learning*, 2024.

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.

Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. Scaling diffusion transformers to 16 billion parameters, 2024. URL https://arxiv.org/abs/2407.11633.

Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*, 2023.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024. URL https://arxiv.org/abs/2403.14608.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021.

Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Hao Kang, and Xin Lu. Infiniteyou: Flexible photo recrafting while preserving your identity. *arXiv preprint arXiv:2503.16418*, 2025.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

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.

Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. Mixlora: Enhancing large language models fine-tuning with lora based mixture of experts. *CoRR*, 2024.

Hongbo Li, Sen Lin, Lingjie Duan, Yingbin Liang, and Ness Shroff. Theory on mixture-of-experts in continual learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7XgKAabsPp.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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*, 2025a.

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*, 2025b.

Zehua Liu, Han Wu, Ruifeng She, Xiaojin Fu, Xiongwei Han, Tao Zhong, and Mingxuan Yuan. Beyond standard moe: Mixture of latent experts for resource-efficient language models. *arXiv e-prints*, pp. arXiv–2503, 2025c.

Fangyuan Mao, Aiming Hao, Jintao Chen, Dongxia Liu, Xiaokun Feng, Jiashu Zhu, Meiqi Wu, Chubin Chen, Jiahong Wu, and Xiangxiang Chu. Omni-effects: Unified and spatially-controllable visual effects generation. *arXiv preprint arXiv:2508.07981*, 2025.

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.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

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

Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *Advances in Neural Information Processing Systems*, 36: 42961–42992, 2023.

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. In *International Conference on Learning Representations*, 2017.

Haotian Sun, Tao Lei, Bowen Zhang, Yanghao Li, Haoshuo Huang, Ruoming Pang, Bo Dai, and Nan Du. Ec-dit: Scaling diffusion transformers with adaptive expert-choice routing. In *The Thirteenth International Conference on Learning Representations*.

Mengyang Sun, Yihao Wang, Tao Feng, Dan Zhang, Yifan Zhu, and Jie Tang. A stronger mixture of low-rank experts for fine-tuning foundation models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=yqyEUcGreT.

Yuhan Wang, Siwei Yang, Bingchen Zhao, Letian Zhang, Qing Liu, Yuyin Zhou, and Cihang Xie. Gpt-image-edit-1.5m: A million-scale, gpt-generated image dataset, 2025. URL https://arxiv.org/abs/2507.21033.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

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*, 2025b.

Qiong Wu, Zhaoxi Ke, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. Routing experts: Learning to route dynamic experts in existing multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*, 2025c.

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*, 2025d.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. In *The Twelfth International Conference on Learning Representations*.

Yike Yuan, Ziyu Wang, Zihao Huang, Defa Zhu, Xun Zhou, Jingyi Yu, and Qiyang Min. Expert race: A flexible routing strategy for scaling diffusion transformer with mixture of experts. In *Forty-second International Conference on Machine Learning*.

Wei Zhang, Jing Liu, and Kai Wang. GEdit: A comprehensive benchmark for general-purpose image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. *arXiv preprint arXiv:2503.07027*, 2025a.

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*, 2025b.

A APPENDIX

A.1 THE USE OF LARGE LANGUAGE MODELS (LLMS)

We acknowledge the use of a large language model (LLM) to aid in the writing process of this manuscript. Its application was confined to language enhancement tasks, such as proofreading for grammatical and spelling errors, rephrasing sentences to improve clarity and readability, and ensuring a consistent formal tone. The LLM served exclusively as a writing assistant.

A.2 TRAINING DATA DETAILS

Our model's versatile capabilities are the direct result of a comprehensive, mixed training dataset, visually summarized in Figure 6. This dataset is meticulously curated to expose the model to a wide spectrum of conditional inputs, combining large-scale, self-synthesized data for complex compositional tasks with established public datasets for foundational spatial control.

Synthesized Data Corpus. To enable sophisticated, instruction-driven editing and composition, we generated a large-scale corpus of training examples. This corpus focuses on tasks that are poorly represented in public datasets but are crucial for real-world applications. Our synthesis pipeline covers the following key areas:

- **Reference-based Generation:** This category involves tasks that condition the output on multiple input images and a textual instruction.
  - **Face Swapping:** Training triplets consist of a source face image, a target person's image, and an instruction to transfer the facial identity.
  - **Multi-Subject Composition:** To teach complex relational reasoning, we generate scenes described by prompts involving multiple, often unrelated, subjects (e.g., "*An old man, a French Bulldog, a tuba, and a rattlesnake in a sunny park*").
  - **Virtual Try-on:** Samples include a person, multiple clothing items, and potentially a pose skeleton, with instructions to dress the person in the specified apparel.
  - **Re-Lighting:** Data consists of a foreground subject and a new background, with the goal of seamlessly integrating the subject into the new lighting environment.
  - **Style Transfer:** We provide a content image and a style reference image (e.g., a photograph and a sketch), instructing the model to render the content in the given artistic style.
- **Single Image Editing:** We employ the **GPT-IMAGE-EDIT-1.5M** dataset (Wang et al., 2025). This corpus contains over 1.5 million high-quality samples, constructed by leveraging GPT-4o to unify and refine existing public datasets such as OmniEdit, HQ-Edit, and UltraEdit. The refinement process ensures superior visual fidelity and semantic alignment for instruction-based editing tasks.

Public Datasets. To build a robust foundation in controllable generation, we incorporate several large-scale, publicly available datasets. These datasets provide strong supervision for fundamental spatial and structural alignment tasks.

- **Spatial Alignment Control:** We leverage datasets that provide dense spatial conditioning maps.
  - **Depth Control:** We utilize datasets such as SubjectSpatial-200K, which provide image-depth map pairs, to train the model to generate scenes with accurate spatial layouts.
  - **Canny Edge Control:** Similarly, we use datasets with image-Canny edge pairs to enable generation from structural outlines.
  - **Pose Control:** We use the widely adopted COCO 2017 dataset (Lin et al., 2014) with its OpenPose keypoint annotations to teach the model to generate human figures conforming to specific pose skeletons.

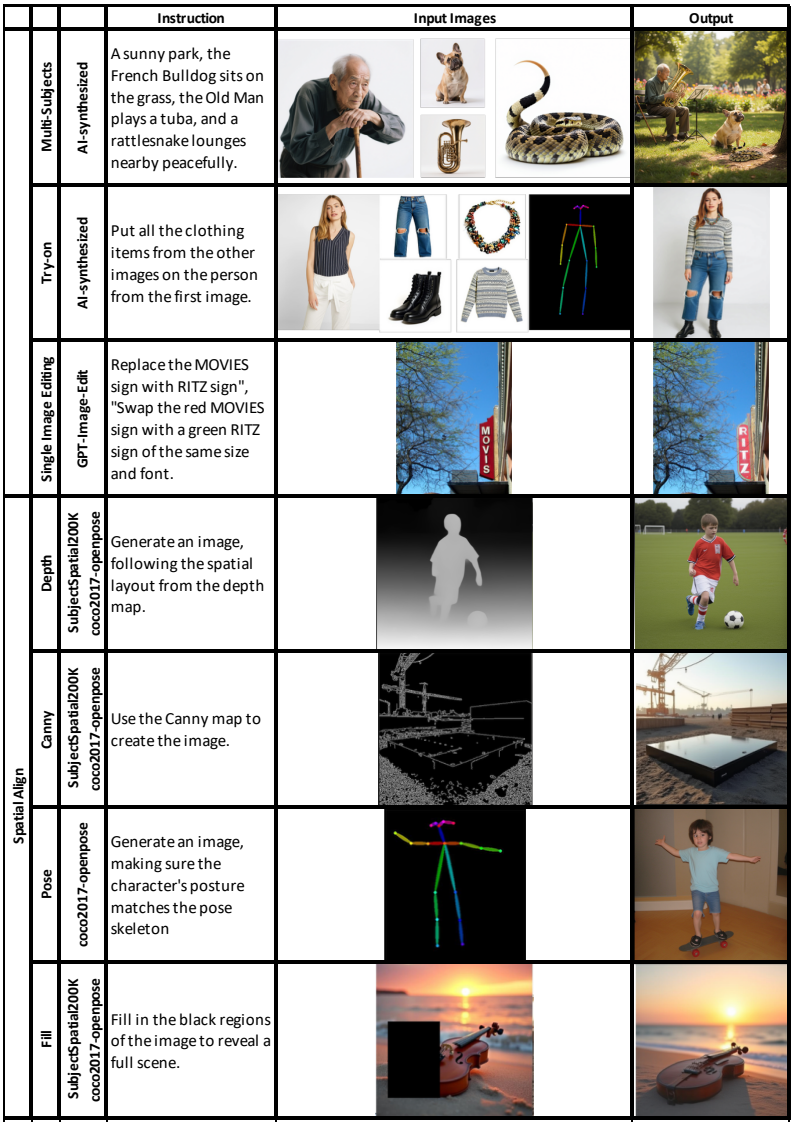


Figure 6: Training Data Details.

- **Image Inpainting (Fill):** Using datasets like SubjectSpatial-200K, we generate training examples by randomly masking regions of an image. This trains the model to fill in missing parts coherently, a crucial skill for image editing and out-painting tasks.

A.3 Ablation Study

Table 6: Ablation study on the MoLE hyper-parameter configuration.

MoLE			Multi-Subject (↑)	Single-Subject (↑)	Depth MSE (↓)	Canny F1 (↑)	Pose F1 (↑)
r=32	N=8	k=2	64.60	6.11	35.25	35.10%	36.47%
r=32	N=8	k=4	**65.68**	**6.15**	**33.34**	**41.51%**	**40.97%**
r=32	N=8	k=6	64.11	6.10	33.35	37.53%	37.90%
r=32	N=8	k=8	65.22	6.05	34.17	37.42%	34.61%
r=64	N=4	k=2	65.06	6.05	33.66	38.27%	34.77%
r=128	N=2	k=1	65.10	**6.15**	35.50	37.52%	34.02%

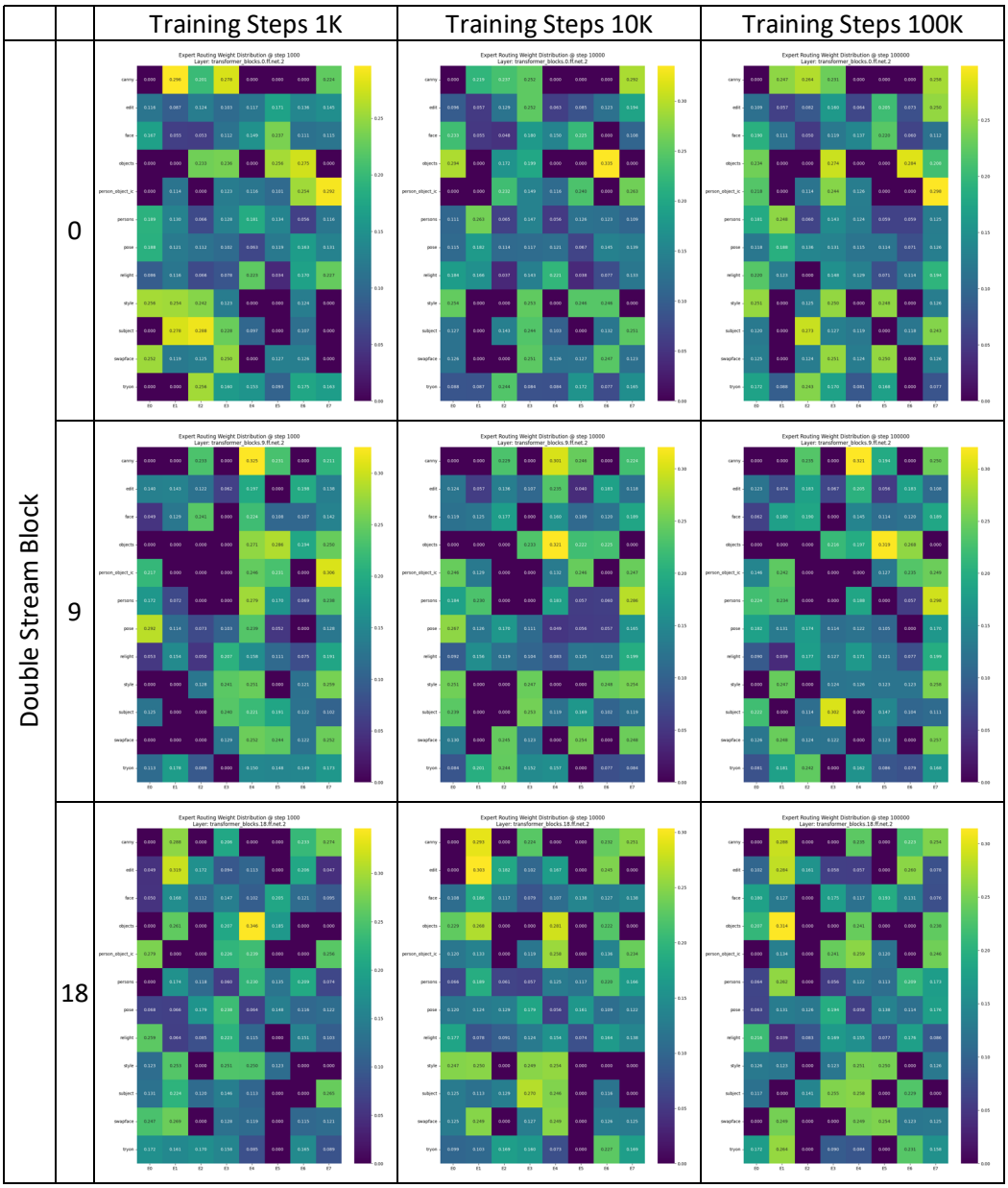


Figure 7: Visualization of the expert routing weight distributions in Double Stream Block. The numbers under the "Single Stream Block" column on the left (0, 9, 18) represent the layer index.

Analysis of MoLE Hyper-parameters. We analyze the MoLE configuration in Table 6 to justify our hyper-parameter choices. The results yield two key insights. First, for a fixed expert pool of $N = 8$ and rank $r = 32$, performance peaks when activating $k = 4$ experts. Activating more experts ($k > 4$), especially in the non-sparse case ($k = 8$), leads to performance degradation, highlighting the importance of sparse activation in mitigating expert interference. Second, for a fixed activation budget ($r \times k \approx 128$), our configuration with a larger pool of diverse, low-rank experts ($N = 8, r = 32$) is more effective than alternatives with fewer, high-capacity experts (e.g., $N = 4, r = 64$). This suggests that a greater diversity of specialized experts is more beneficial than individual expert capacity. These findings validate our use of the ($r = 32, N = 8, k = 4$) configuration for all main experiments.

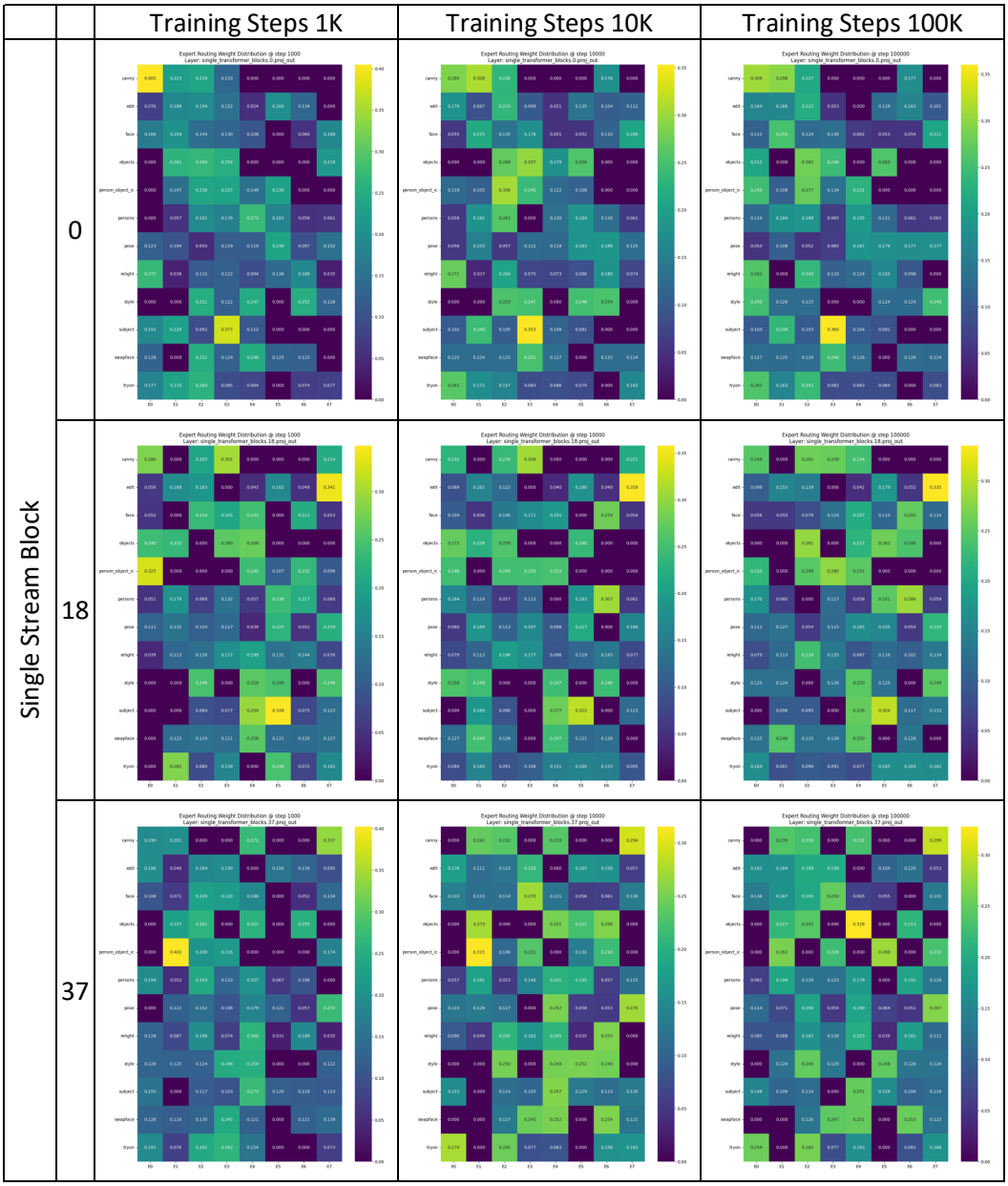


Figure 8: Visualization of the expert routing weight distributions in Single Stream Block. The numbers under the "Single Stream Block" column on the left (0, 18, 37) represent the layer index.

Analysis of Expert Specialization Dynamics. To investigate the learning dynamics of our MoE framework, we visualize the expert routing weight distributions at different training stages (1K, 10K, and 100K steps) and across various model depths, as shown in Figure 7 and Figure 8 The analysis reveals two key findings.

First, we observe a clear evolutionary trajectory from an initial, diffuse routing policy to a stable and highly specialized one. At 1K steps, the weights are distributed relatively evenly, indicating an early exploratory phase. By 100K steps, the distributions become notably sparse and peaked, with specific experts consistently chosen for particular task categories. This progression from a generalized to a specialized routing policy demonstrates the effective convergence of our training objectives in guiding functional disentanglement.

Second, the nature of expert specialization varies with model depth, suggesting a hierarchical division of labor. In early layers (e.g., Single Stream Block 0), experts tend to specialize in processing fundamental input types, such as distinguishing "subject" from "objects". In medial layers (e.g., Single Stream Block 18), we observe the emergence of clear, task-level specialists, with distinct experts strongly favoring tasks like "canny" or "swapface". In later layers (e.g., Single Stream Block 37), the routing pattern often becomes more distributed again, suggesting a shift from semantic task execution to a more collaborative final synthesis stage involving multiple experts.

A.4 QUALITATIVE COMPARISON

Figure 9: Qualitative comparison on OmniContext benchmark.

Figure 10: Qualitative comparison on Xverse benchmark.

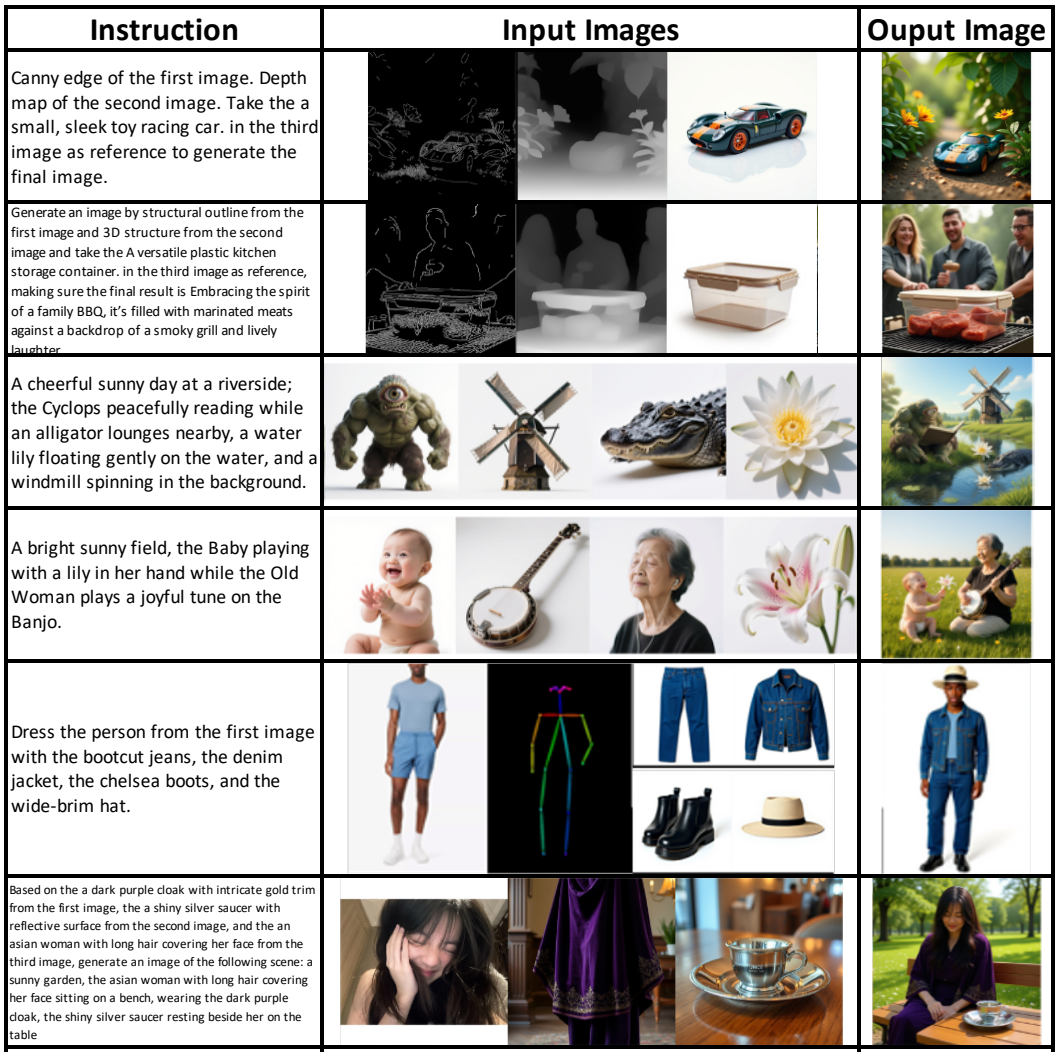


Figure 11: Sample outputs generated by our model (InstructMoLE) from diverse, multi-modal instructions.

A.5 Qualitative Analysis of Extreme Scenarios

To rigorously assess the operational boundaries of our method, we conducted stress tests on extreme scenarios involving textual noise, implicit reasoning, and high-complexity composition, as visualized in Figure 12.

Limitations and Common Failure Modes. Our analysis reveals a discernible threshold for robustness. In Case #1, characterized by severe typos (e.g., "Chnage the bckgrnd..."), both InstructMoLE and the baseline fail to resolve the semantic intent. This indicates that while the CLIP semantic anchor provides stability, it cannot fully compensate for T5 token embeddings when textual corruption is excessive. Similarly, in the ultra-complex Case #5, which requires simultaneous reference binding for three distinct entities ("robot," "boy," "girl"), both models struggle to correctly map identities to the generated subjects. This points to a capacity bottleneck in the underlying cross-attention mechanism of the foundation model when handling high reference loads, a limitation that persists regardless of the routing strategy.

Superiority in Reasoning and Coherence. Despite these boundary conditions, InstructMoLE exhibits significantly stronger capabilities in *implicit reasoning* and *atmospheric consistency*. In Case #2 ("Make it look dangerous"), our model successfully modulates the robot's expression and lowers

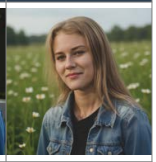
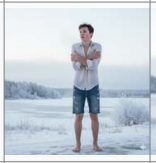
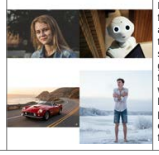
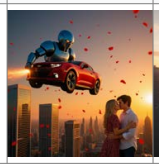
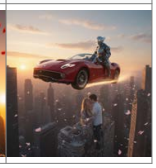
#	Input Image	Instruction	Test Objective	Flux.1 Kontext (dev)	Ours
1		Chnage the bckgrnd to a bueatyful bceah at susnet.	Verify if the model can ignore severe typos and correctly identify semantics for 'background', 'beach', and 'sunset' via the CLIP anchor.		
2		Make it look dangerous.	Test the model's ability to perform abstract semantic editing based on CLIP associations (e.g., lighting, mood) without explicit object references.		
3		Replace the red car with a vintage blue truck and change the season to snowy winter.	Test the router's ability to handle simultaneous local object manipulation and global style/environment changes without feature leakage (e.g., ensuring the truck doesn't become snowy white or the snow doesn't turn blue).		
4		The person in the photo is cold, give them something to wear.	Test logical reasoning capabilities. The model must infer the visual implication of 'cold' (needs clothing) and generate an appropriate object (jacket/scarf) that fits the subject's pose, rather than just changing the color temperature.		
5		Referring to the given image, generate the following scene: A futuristic robot with metallic blue armor is driving a sleek red sports car, soaring through the air above a bustling cityscape at sunset. The car leaves a trail of golden sparks as it glides between skyscrapers. Meanwhile, in the foreground of the image, a young couple—a boy with dark hair wearing a white shirt and a girl with long blonde hair in a floral dress—are sharing a passionate kiss under the warm glow of the evening light, with cherry blossom petals gently falling around them.	Test the model's ability to: (1) handle multiple independent objects simultaneously without attribute leakage (e.g., robot's blue color doesn't affect the car's red color, car's aerial position doesn't affect the couple's ground position); (2) understand and maintain spatial relationships (aerial flying car vs. foreground kissing couple); (3) combine complex actions and scenes (driving/flying vs. kissing) in a coherent composition; (4) preserve distinct visual attributes for each object while creating a unified scene.		

Figure 12: **Extreme Scenario Stress Test: Flux.1 Kontext vs. InstructMoLE.** We evaluate robustness across five edge cases spanning input noise, implicit reasoning, and complex composition. While both models share limitations under severe textual corruption (#1) and high reference load (#5), InstructMoLE demonstrates superior capabilities in abstract atmospheric inference (#2) and logical causality (#4), successfully executing implicit edits where the baseline fails.

scene lighting to align with the semantic tone, whereas the baseline output remains incongruously bright. Crucially, in Case #4 ("The person is cold"), InstructMoLE correctly infers the causal implication to generate clothing (a jacket), a logical step the baseline fails to execute. Furthermore, in multi-attribute editing (Case #3), our method achieves superior spatial and lighting integration of the inserted object.