WHEN UNIFIED MODELS MEET VISUAL EXPERTS: LIGHTWEIGHT COLLABORATION VIA GENERATIVE HANDOFF

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models (UMMs) offer a single, flexible interface for visual understanding, generation, and editing, but specialized image generators and editors often achieve better visual quality on the tasks they are built for. Getting the two to work together, however, typically requires additional alignment training, which is costly and can compromise the UMM's broader capabilities. We find that, for several UMM–expert pairs, their latent spaces and generative trajectories are already compatible enough for one model to continue sampling from a state produced by the other. Therefore, we introduce **Handoff**, a training-free collaboration strategy, where the UMM performs the early sampling steps, then passes its intermediate latent state directly to a visual expert, which completes generation from a selected timestep. Handoff retains the UMM's instruction-following and semantic capabilities while improving perceptual quality, without learned connectors or cross-model alignment. We further introduce **Partial On-Policy Distillation (P-OPD)** for settings where the expert is unavailable at inference time. P-OPD distills the expert's behavior only over the later sampling steps, using a timestep-gated LoRA, so that the resulting UMM can reproduce the benefit without calling the expert. Across image generation and editing tasks and multiple UMM–expert pairs, both methods consistently improve the underlying UMM. Handoff can even outperform both standalone models, including on the challenging LongText-Bench and CVTG-2K benchmarks. Together, these results show that UMMs and specialized visual experts can collaborate through their existing generative representations, with little or no additional alignment.

1 INTRODUCTION

Visual generation is advancing along two complementary directions. Unified multimodal models (UMMs) (Deng et al., 2025; Chen et al., 2025b; Wu et al., 2025b; Tian et al., 2026; Diao et al., 2026) provide a flexible interface for understanding, generation, and editing, while specialized image generators and editors (Esser et al., 2024; Black Forest Labs, 2024; Labs et al., 2025; Wu et al., 2025a), which we call *experts*, often achieve better perceptual quality. Their strengths are complementary (Fig. 1): a UMM may correctly decide what content belongs where but render it poorly, whereas an expert may produce a cleaner image while missing parts of the instruction. Ideally, the UMM should determine *what* to generate and the expert should improve *how* it is rendered.

Existing ways of combining them, however, either require substantial training or discard much of what the UMM has already decided. Learned connectors (Koh et al., 2023; Wu et al., 2024b; Pan et al., 2025; Zeng et al., 2026) can pass rich information between models, but require pairwise alignment training and couple the two models together. Tool-use pipelines (Shen et al., 2023; Wang et al., 2024b) avoid training, but typically communicate through only a prompt or a finished image. The expert must then infer again decisions such as layout, object placement, or which text belongs to which object. **Can we connect the two models without training a new interface, while preserving the decisions already made during generation?**

We find that, for several independently developed UMM–expert pairs, the answer is yes. Their generation processes operate in compatible latent spaces and follow similar enough dynamics that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

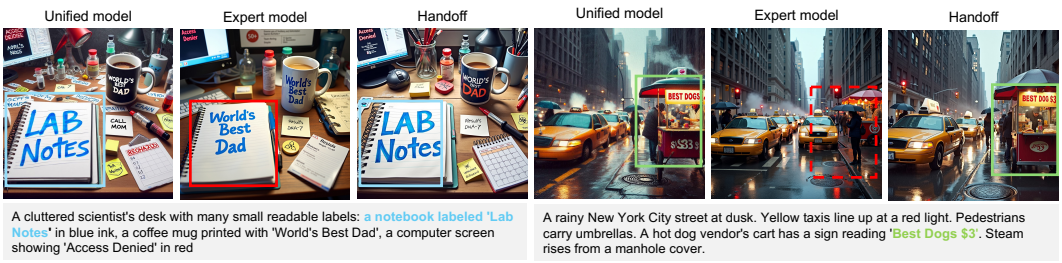

Figure 1: **Comparison of unified model (Bagel) and expert models (Flux.1.dev).** BAGEL captures key requested content but exhibits rendering artifacts, while FLUX.1-dev produces cleaner images yet misassigns or omits requested text. Handoff preserves key content from BAGEL while improving text rendering and visual details.

an expert can continue sampling from an intermediate state produced by a UMM. We call this intermediate latent a *generative state*. Although switching denoisers partway through sampling has mainly been studied within one model family (Balaji et al., 2023; Podell et al., 2023; Pan et al., 2024; Cheng et al., 2025), we find that it can also work across separately trained models: **an expert can pick up a UMM's generation midway through sampling while preserving the semantic decisions already encoded in the state.**

Based on this observation, we introduce *Handoff*, a training-free collaboration strategy (Fig. 2). The UMM performs the early sampling steps and passes its intermediate state to the expert, which completes the remaining steps under its own conditioning. Since early steps largely determine content and layout while later steps refine appearance (Balaji et al., 2023; Meng et al., 2022), this division lets the UMM establish the semantics before the expert improves the final rendering. Handoff requires no learned connector, alignment training, or additional sampling steps. For editing, the expert can even be a text-to-image model with no editing interface of its own, since the source content and edit decisions are already carried by the generative state.

Handoff still requires both models at inference time. To internalize the expert instead, we introduce *Partial On-Policy Distillation* (P-OPD). Naive distillation over the full trajectory can overwrite the early steps where the UMM interprets the instruction, while distillation on pre-collected trajectories trains on off-policy states. P-OPD therefore distills only the later part of the trajectory: a timestep-gated LoRA is trained to match the expert's velocity on the UMM's own rollouts and is activated only after the corresponding handoff timestep.

We evaluate Handoff and P-OPD with three UMMs and four experts from two VAE families, across image generation and editing. Handoff consistently improves the underlying UMMs and can outperform both standalone models. For example, BAGEL→FLUX raises LongText-Bench from 0.390 to 0.655 (FLUX: 0.612) and CVTG-2K from 0.356 to 0.596 (FLUX: 0.497). For editing, BAGEL→FLUX.1-Kontext raises GEdit-Bench from 6.83 to 7.04, above FLUX.1-Kontext alone at 6.57. Without the expert at inference time, P-OPD improves BAGEL on all six benchmarks, including CVTG-2K from 0.356 to 0.543. These results suggest that intermediate generative states can serve as a practical interface between independently developed visual models. Rather than aligning entire models, we can let each model contribute where it is strongest, either by handing off the generation directly or by distilling only the part of the trajectory that benefits from the expert.

2 Related Work

Unified Multimodal Models Unified multimodal models aim to support visual understanding, generation, and editing within one model. Early approaches extend autoregressive next-token prediction to discrete visual representations, as exemplified by Chameleon and Janus Team (2025); Wang et al. (2024a); Wu et al. (2024a); Chen et al. (2025b); Wu et al. (2025d;c), while other works explore discrete diffusion within unified transformers Xie et al. (2025a). Recent models increasingly adopt continuous generative modeling: Transfusion jointly models discrete text and continuous images, while BAGEL and related approaches combine autoregressive language modeling with

diffusion or flow matching Zhou et al. (2024); Ma et al. (2025); Shi et al. (2025); Deng et al. (2025); Xie et al. (2025b). Other architectures condition dedicated diffusion or flow-based generators on multimodal representations Ge et al. (2025); Tong et al. (2024); Pan et al. (2025); Chen et al. (2025a); Peng et al. (2026). More recent systems such as OmniGen2 further expand the unified interface to richer instruction, reference-image, and interleaved multimodal contexts Wu et al. (2025b); Tian et al. (2026). Despite their broad task coverage, unified models do not consistently match specialized generators and editors on challenging generation and manipulation tasks, including long-text rendering and precise image editing Geng et al. (2025); Tai et al. (2026); Chang et al. (2025b); Liu et al. (2025a).

Image Generation and Editing Models Specialized image generators have achieved strong visual quality and text alignment through diffusion and flow-based architectures Rombach et al. (2022); Peebles & Xie (2023); Podell et al. (2023); Esser et al. (2024); Black Forest Labs (2024); Xie et al. (2024); Gao et al. (2025); Wu et al. (2025a). For text-rich generation, methods such as TextDiffuser and AnyText improve visual text rendering through glyph-aware conditioning, specialized text representations, and text-rich training data Chen et al. (2023); Tuo et al. (2024); Wang et al. (2025). Image editing has evolved from inversion and attention-based control Meng et al. (2022); Hertz et al. (2022); Mokady et al. (2022) toward instruction-conditioned models such as InstructPix2Pix and large-scale editing systems trained on diverse image-editing pairs Brooks et al. (2023); Zhang et al. (2024a); Sheynin et al. (2023). Recent models such as FLUX.1-KontextLabs et al. (2025) and Qwen-Image-EditWu et al. (2025a) further support high-fidelity reference-conditioned editing and complex text manipulation. Although these models excel at specific tasks, they support a narrower range of tasks than unified multimodal models.

Cross-Model Collaboration. A growing body of work explores cross-model collaboration. GILL and NExT-GPT learn modules to connect language models with generators Koh et al. (2023); Wu et al. (2024b), while other approaches guide generation through layout or regional planning Feng et al. (2023); Lian et al. (2024); Yang et al. (2024). Collaboration also occurs during denoising: eDiff-I uses noise-specific denoisers Balaji et al. (2023), RealCompo combines differently conditioned predictions Zhang et al. (2024b), and T-Stitch switches denoisers along a shared trajectory without additional training Pan et al. (2024). However, these approaches require additional training, rely on explicit conditioning plans, or focus on collaboration among task-specific generators. In contrast, we enhance unified models through training-free collaboration with image experts, bridging different architectures and conditioning interfaces through intermediate states in a shared latent space.

3 METHOD

Figure 2 provides an overview of our two strategies. Handoff transfers an intermediate generative state from the unified model to an expert, which completes the remaining sampling steps. P-OPD instead distills the expert's suffix predictions into the unified model.

3.1 PRELIMINARIES

Unified Model. A unified multimodal model $\mathcal{M}_U$ supports visual understanding and generation within a single framework. We consider its flow-based image generation process. Given a native condition c_U, initial noise $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and a schedule $0 = t_0 < \cdots < t_K = 1$, sampling follows

$$\mathbf{z}_{k+1} = \mathbf{z}_k + \Delta t_k \, \mathbf{v}_U(\mathbf{z}_k, t_k; c_U), \tag{1}$$

where $\Delta t_k = t_{k+1} - t_k$ and $\mathbf{v}_U$ is the predicted velocity. Each intermediate latent $\mathbf{z}_k$ is called a *generative state*. The final image is decoded as $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z}_K)$.

Expert Model. An expert $\mathcal{M}_E$ specializes in a visual generation task and predicts a velocity $\mathbf{v}_E(\mathbf{z}_k, t_k; c_E)$ under its condition c_E. We consider experts that share the unified model's VAE and use compatible noise and time setting, allowing both models to operate on the same generative states.

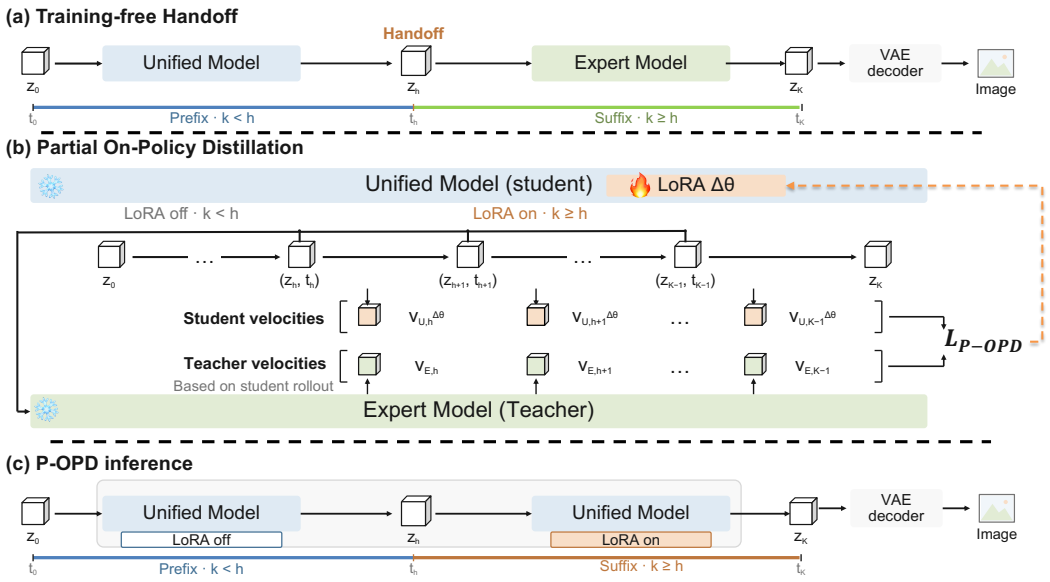

Figure 2: **Overview of Handoff and P-OPD.**(a) Training-free Handoff: the unified model performs the first h sampling steps and transfers its intermediate state to the expert, which completes the remaining $K - h$ steps. Both models remain frozen. (b) P-OPD training: the unified-model student generates fresh on-policy rollouts and matches its suffix velocities to targets from the frozen expert at the same states and timesteps. Only the student's timestep-gated LoRA adapters are updated. (c) P-OPD inference: the unified model runs the full trajectory, with LoRA enabled only for $k \geq h$;

3.2 Training-Free Handoff

Handoff transfers a generative state from the unified model to the expert. The unified model performs the first h sampling steps; the expert then receives $\mathbf{z}_h$ at timestep t_h and completes the others:

$$\mathbf{z}_{k+1} = \mathbf{z}_k + \Delta t_k \begin{cases} \mathbf{v}_U(\mathbf{z}_k, t_k; c_U), & k < h, \\ \mathbf{v}_E(\mathbf{z}_k, t_k; c_E), & k \geq h. \end{cases} \tag{2}$$

For text-to-image generation, c_U and c_E encode the same prompt through each model's native conditioning module. The expert continues on the unified model's remaining timestep grid $\{t_h, \ldots, t_K\}$. Both models remain frozen, and handoff requires no connector or additional training.

3.3 Training-Based Distillation

We introduce *Partial On-Policy Distillation* (P-OPD) to transfer the expert's suffix predictions into the unified model. The expert model serves as teacher, and unified model serves as student.

We freeze the student model and add trainable LoRA parameters $\Delta\theta$ to its generation module. At each training iteration, the student generates a fresh rollout, with LoRA disabled for the first h steps and enabled afterwards. At every suffix state, the frozen teacher provides a velocity target at the same latent and timestep, and all $K - h$ suffix states are supervised:

$$\mathcal{L}_{\text{P-OPD}} = \mathbb{E}_{p,\mathbf{z}_0}\left[\frac{1}{K-h}\sum_{k=h}^{K-1}\left\|\mathbf{v}_U^{\Delta\theta}(\mathbf{z}_k, t_k; c_U) - \mathbf{v}_E(\mathbf{z}_k, t_k; c_E)\right\|_2^2\right], \tag{3}$$

where $\mathbf{v}_U^{\Delta\theta}$ denotes the velocity with LoRA enabled. Only $\Delta\theta$ is updated. The distillation is on-policy because states come from the current student, and partial because supervision and adaptation are restricted to the suffix. Setting $h = 0$ recovers global on-policy distillation, which we compare against in Sec. 4.3. At inference, the unified model alone runs the full trajectory, with LoRA enabled only for $k \geq h$; no expert is required.

3.4 Extension to Image Editing

Both methods extend to image editing. The UMM takes a source image and an editing instruction as c_U. An editing expert uses both inputs as c_E, whereas a text-to-image expert uses only the instruction, receiving source-image information solely through $\mathbf{z}_h$ and enabling editing without editing-specific training. Handoff otherwise follows Eq. (2). P-OPD uses source-image–instruction pairs for student rollouts and an editing expert as the teacher in Eq. (3).

4 Experiments

4.1 Experimental Settings

Model pairs. We evaluate both image generation and image editing in two model families. In the FLUX-VAE family, BAGEL (Deng et al., 2025) and OmniGen2 (Wu et al., 2025b) serve as unified models, paired with FLUX.1-dev (Black Forest Labs, 2024) for generation and FLUX.1-Kontext (Labs et al., 2025) for editing. In the Qwen-Image-VAE family, InternVL-U (Tian et al., 2026) serves as the unified model, paired with Qwen-Image (Wu et al., 2025a) for generation and Qwen-Image-Edit for editing.

Training-free settings. For BAGEL and OmniGen2, we use $K = 50$ sampling steps with a timestep shift of 3.0, handing off at $h = 10$ for generation and $h = 30$ for editing, for both editing and text-to-image experts. InternVL-U follows its default $K = 20$ schedule, with $h = 4$ for generation, $h = 9$ for editing, and $h = 15$ for handoff to the text-to-image expert.

Training-based settings. We train separate rank-256 LoRA adapters for BAGEL generation and editing, using frozen FLUX.1-dev and FLUX.1-Kontext [dev] experts, respectively; the BAGEL backbone also remains frozen. Both tasks use the same $K = 50$ schedule with timestep shift 3.0 as in the training-free setting, activating the adapters at $h = 10$ and $h = 30$, respectively. We train each adapter for 1,200 AdamW updates, with a learning rate of 10^{-4}, zero weight decay, and gradient clipping at 1.0. Training uses 1,600 synthetic prompts for generation and 1,600 source-image–instruction pairs sampled from the MagicBrush training split (Zhang et al., 2024a) for editing.

Evaluation benchmarks. For image generation, we use LongText-Bench-EN (Geng et al., 2025) and CVTG-2K (Tai et al., 2026) to evaluate text rendering, and OneIG-Bench (Chang et al., 2025a) and WISE (Niu et al., 2025) to assess general generation capabilities. We report Text Score, Word Accuracy, the official OneIG category and overall scores, and WISE Verified, respectively. For image editing, we evaluate on GEdit-Bench-EN (Liu et al., 2025b) and ImgEdit (Ye et al., 2025) to assess editing instruction following and visual quality. We report Semantic Consistency (SC), Perceptual Quality (PQ), and Overall (O) on GEdit-Bench-EN, and the official Overall score on ImgEdit. We follow the official evaluation protocols for all benchmarks.

4.2 Main Results

Generation. Table 1 shows that Handoff improves LongText-Bench, CVTG-2K, and OneIG-Bench overall for all three unified models. BAGEL improves from 0.390 to 0.655 on LongText-Bench and from 0.356 to 0.596 on CVTG-2K, exceeding standalone FLUX on both. For OmniGen2, Handoff reaches 0.590 on CVTG-2K, exceeding both OmniGen2 (0.516) and FLUX (0.497), even though the expert has the lower standalone score. InternVL-U also exceeds both constituent models on OneIG-Bench overall, reaching 0.55 compared with 0.50 for InternVL-U and 0.54 for Qwen-Image. These results support complementary strengths between the unified models and experts.

P-OPD improves standalone BAGEL on all four benchmarks. It raises CVTG-2K word accuracy from 0.356 to 0.543 and OneIG-Bench overall from 0.41 to 0.46, exceeding standalone FLUX on both. It also achieves higher OneIG-Bench Text and overall scores than Handoff, while retaining BAGEL's Alignment score of 0.82. These gains are obtained without an expert at inference, showing that P-OPD can transfer part of the expert's generation capability to the unified model.

Table 1: **Comparison of image generation performance across four benchmarks.** Gray entries denote standalone experts, and blue rows highlight our methods. Bold indicates the best result within each unified-model block. Higher scores are better for all metrics.

Model	Method	LongText	CVTG-2K	OneIG-Bench						WISE
		Text Score	Word Acc.	Align	Text	Reas.	Div.	Style	Overall	Verified
FLUX-VAE family										
FLUX	*Expert*	0.612	0.497	0.80	0.58	0.25	0.24	0.37	0.45	0.45
BAGEL	Baseline	0.390	0.356	**0.82**	0.41	0.19	**0.24**	0.39	0.41	0.35
	Handoff	**0.655**	**0.596**	0.80	0.60	**0.25**	**0.24**	0.38	0.45	**0.42**
	P-OPD	0.456	0.543	**0.82**	**0.65**	0.20	**0.24**	**0.41**	**0.46**	0.40
OmniGen2	Baseline	0.561	0.516	0.80	0.68	**0.27**	0.24	0.38	0.47	**0.34**
	Handoff	**0.647**	**0.590**	**0.81**	**0.75**	**0.27**	**0.26**	**0.40**	**0.50**	**0.34**
Qwen-Image-VAE family										
Qwen-Image	*Expert*	0.925	0.829	0.88	0.99	0.28	0.19	0.39	0.54	0.51
InternVL-U	Baseline	0.541	0.623	0.82	0.74	0.27	**0.25**	0.40	0.50	0.46
	Handoff	**0.880**	**0.732**	**0.87**	**0.90**	**0.29**	0.24	**0.43**	**0.55**	**0.47**

Table 2: **Image editing results on GEdit-Bench-en and ImgEdit.** † denotes handoff to a image generation expert rather than a dedicated image-editing expert. Bold marks the highest score per metric within each unified-model block.

Model	Method	GEdit-Bench-en			ImgEdit									
		SC ↑	PQ ↑	O ↑	Add ↑	Adj. ↑	Extr. ↑	Repl. ↑	Rem. ↑	BG ↑	Style ↑	Hybrid ↑	Action ↑	Overall ↑
FLUX-VAE family														
FLUX.1-Kontext [dev]	*Expert*	7.02	7.58	6.57	4.14	4.20	2.11	4.32	3.95	3.94	4.46	2.43	4.21	3.79
BAGEL	Baseline	7.68	6.71	6.83	3.51	3.64	**1.82**	3.12	2.68	3.38	4.07	2.52	4.09	3.17
	Handoff	**7.78**	7.13	**7.04**	4.00	3.93	1.69	3.51	2.92	3.52	4.20	2.71	4.43	3.38
	Handoff†	7.20	**7.20**	6.78	**4.40**	**4.10**	**1.82**	**3.69**	**3.67**	**3.71**	**4.50**	**3.29**	**4.70**	**3.68**
	P-OPD	7.73	7.00	6.97	4.08	3.67	1.78	3.50	2.94	3.50	3.81	3.05	4.09	3.31
OmniGen2	Baseline	6.84	7.40	6.41	3.67	3.76	1.91	3.59	2.82	3.87	4.53	**3.05**	4.79	3.47
	Handoff	**6.93**	7.50	6.49	3.88	**4.02**	2.21	**4.07**	3.60	4.17	**4.70**	2.90	4.39	3.77
	Handoff†	6.87	**7.75**	**6.62**	**4.03**	3.98	**2.29**	4.02	**4.06**	**4.24**	4.24	**3.05**	**4.91**	**3.85**
Qwen-Image-VAE family														
Qwen-Image-Edit	*Expert*	8.16	7.71	7.66	4.55	4.43	4.35	4.69	4.64	4.49	4.66	3.52	4.55	4.51
InternVL-U	Baseline	7.56	6.40	6.66	3.77	3.37	3.48	**3.63**	3.97	3.48	4.06	**3.24**	3.88	3.67
	Handoff	**7.97**	**7.54**	**7.40**	**4.15**	**4.10**	**3.61**	3.59	**4.44**	**3.89**	**4.44**	**3.24**	**4.42**	**4.01**

Editing. Table 2 shows that handoff to editing experts improves the overall scores on both GEdit-Bench-en and ImgEdit for all three unified models. InternVL-U, for example, improves from 6.66 to 7.40 on GEdit-Bench-en and from 3.67 to 4.01 on ImgEdit.

Using image generation expert yields higher ImgEdit scores than using FLUX.1-Kontext for both BAGEL (3.68 vs. 3.38) and OmniGen2 (3.85 vs. 3.77), despite FLUX.1-dev receiving no source image. However, the GEdit-Bench-en results are mixed. For BAGEL, perceptual quality improves from 6.71 to 7.20, but semantic consistency decreases from 7.68 to 7.20, and the overall score falls from 6.83 to 6.78. OmniGen2 instead improves its overall score from 6.41 to 6.62. These results show that a generation expert can continue an editing trajectory, but improvements in visual quality do not always coincide with better semantic consistency.

P-OPD improves BAGEL from 6.83 to 6.97 on GEdit-Bench-en and from 3.17 to 3.31 on ImgEdit. These gains are obtained without an expert at inference, showing that P-OPD can transfer part of the expert's editing capability to the unified model.

Qualitative results. Figure 3 illustrates how Handoff preserves the unified model's scene structure while refining content and visual appearance. In the cafe example, it retains BAGEL's multi-line signage and composition while producing more regular lettering. In the Monument Valley and market examples, it preserves the main scene layout while improving landscape detail and visual realism, respectively. P-OPD exhibits similar improvements without an expert at inference: it better covers the requested signage content in the street-cart example, although spelling errors remain, and improves

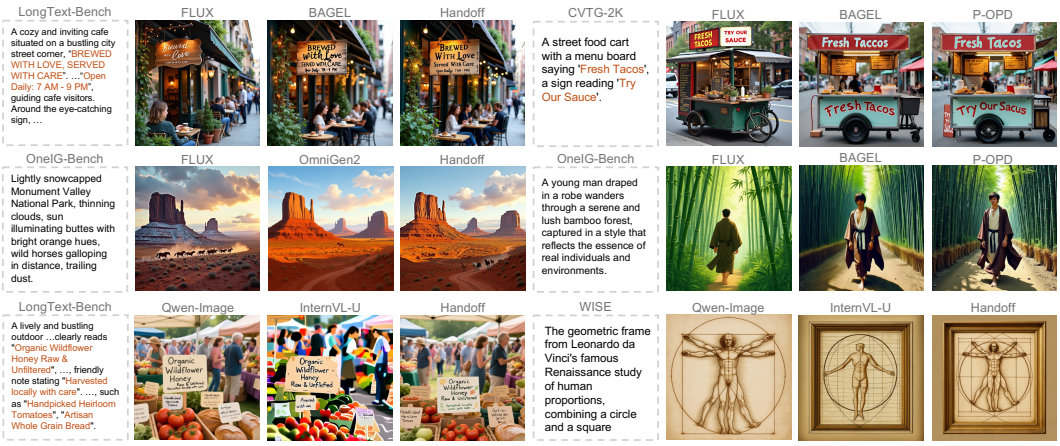

Figure 3: **Qualitative results on image generation.** Benchmark names indicate the source of each prompt, and orange text highlights the target text to be rendered.

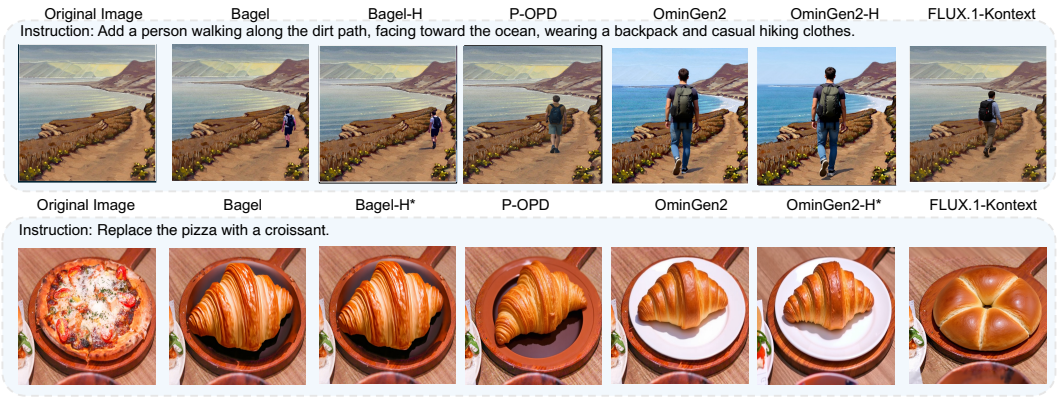

Figure 4: **Qualitative results on image editing.** The suffixes -H and -H* denote handoff to an image-editing expert and a text-to-image generation expert, respectively.

visual fidelity in the bamboo example. In the WISE example, Handoff produces the characteristic multi-pose human figure while retaining InternVL-U's decorative frame. This suggests that expert continuation can revise concept-related content while remaining influenced by the scene structure established in the prefix. Figure 4 presents qualitative editing results for Handoff and P-OPD. In the croissant example, standalone FLUX.1-Kontext produces a rounded bun, whereas handoff to the generation expert retains the unified model's correct replacement and refines the pastry texture. P-OPD also performs the requested edits with BAGEL alone at inference.

4.3 Ablation Studies

Timestep selection. We vary the handoff step h for BAGEL→FLUX.1-dev while fixing the total sampling budget at $K = 50$. As shown in Fig. 6, performance generally improves as the handoff is delayed beyond the earliest steps, but decreases with later handoffs. This suggests that neither immediate switching nor a short expert continuation yields the best performance in this setting.

Figure 5 illustrates how the handoff position affects the generated images. In these examples, later-handoff outputs more closely resemble standalone BAGEL in overall composition and object arrangement, while earlier handoffs allow larger changes in the final appearance. Thus, the handoff position affects both generation quality and the extent to which the output retains image structure.

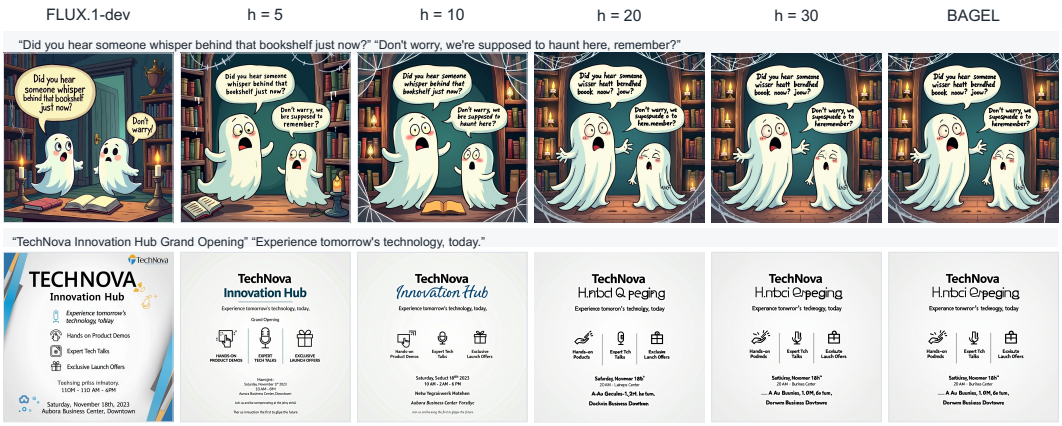

Figure 5: **Qualitative comparison of different handoff steps.** Target text excerpts are provided above each row.

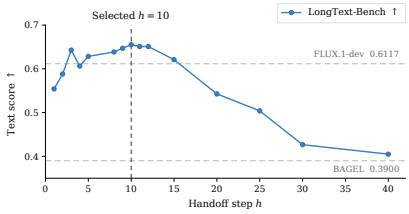

Figure 6: Effect of the handoff step h on LongText-Bench.

Table 3: **Ablation of rollout policy and adaptation scope in P-OPD.** Bold indicates the best result among the four training variants.

Variant	On-policy	Partial	LongText	CVTG-2K	OneIG	WISE
Baseline	–	–	0.390	0.356	0.41	0.35
Global + Offline	✗	✗	0.355	0.427	0.44	**0.41**
Global + On-policy	✓	✗	0.362	0.452	0.43	**0.41**
Partial + Offline	✗	✓	0.385	0.461	0.44	0.39
P-OPD	✓	✓	**0.456**	**0.543**	**0.46**	0.40
Handoff	–	–	0.655	0.596	0.45	0.42

Training Method. We compare P-OPD with three alternative training configurations that vary the adaptation scope and rollout policy. As shown in Table 3, P-OPD is the only trained variant that improves BAGEL on all four benchmarks and achieves the best LongText-Bench, CVTG-2K, and OneIG-Bench overall scores among these configurations. For example, P-OPD reaches 0.543 on CVTG-2K, compared with 0.452 for global on-policy training and 0.461 for partial offline training. The alternative configurations improve several benchmarks but fail to improve LongText-Bench over the baseline. Although global adaptation performs slightly better on WISE, P-OPD provides more consistent improvements across the evaluated benchmarks, supporting the effectiveness.

4.4 Analysis

Text rendering analysis. Handoff yields particularly strong gains on text rendering (Table 1). As illustrated in the ghost-dialogue example in the first row of Figure. 5 illustrates complementary limitations: FLUX alone may allocate unsuitable regions for the requested text, while the unified model provides a more appropriate layout but renders some words inaccurately. These examples suggest that handoff supplies FLUX with useful spatial organization from the unified-model prefix, allowing its text-rendering capability to be used more effectively. This layout–rendering complementarity offers a possible explanation for why the composed trajectory can outperform either model alone.

Editing through a generation expert. We analyze why a generation expert can continue an editing task despite lacking a native editing interface (Fig. 7). All variants start from the same BAGEL state at $h = 30$. Without this state, FLUX.1-dev follows only the textual semantics of the instruction and ignores the source image (*Prompt only*). Given the state, the source content, layout, and placement of the edited text are preserved even with an empty prompt (*State only*), indicating that the structure of the edit has already been committed to the latent. The prompt nevertheless matters for fine-grained details: BAGEL misspells the target text ("Lomely Bear"), which FLUX.1-dev corrects when conditioned on the instruction ("Lonely Bear") but not with an empty prompt ("L'omely Bear"). Interestingly, the editing expert FLUX.1-Kontext, conditioned on both the instruction and the

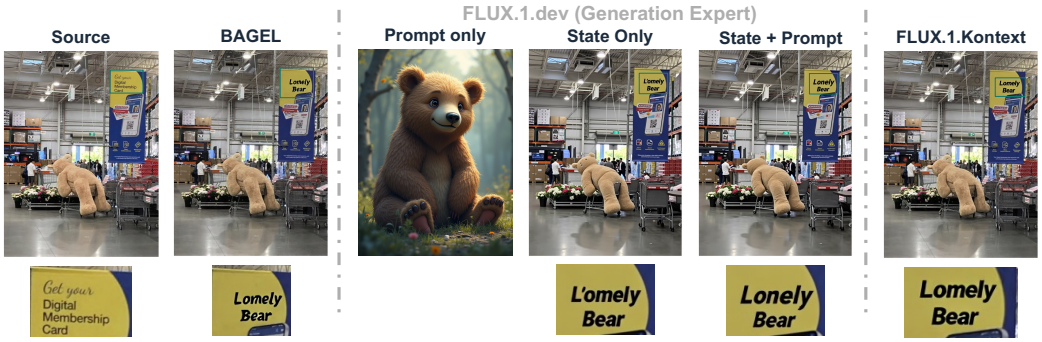

Figure 7: **Illustrative examples of editing through a generation expert** (FLUX.1-dev). The editing instruction is "Change the text in the picture to 'Lonely Bear'."

reference image, retains BAGEL's misspelling, suggesting that it more conservatively preserves the handoff state. We hypothesize that this difference underlies the trade-off in Table 2: by re-synthesizing more aggressively, a generation expert can correct prefix errors and improve perceptual quality, but may also drift from source details, as reflected in the lower semantic consistency of BAGEL-H†. Overall, the unified model establishes the structure of the edit, while the expert refines appearance and resolves details that remain undetermined at handoff.

State influence. We use FLUX.1-dev as the receiver, providing the target instruction and transmitted state but no source image. We fix the instruction, handoff timestep, and sampling configuration, varying only the state. On LongText-Bench shuffling reduces target recall from 0.655 to 0.561 and increases donor-only recall from 0.014 to 0.037. Reconstruction and re-noising achieve a target recall of 0.629, above standalone FLUX (0.612) but below direct handoff (0.655), while norm-matched random states yield only 0.008. Detailed results and further discussion are provided in the appendix A.3.

5 Discussion and Limitations

Discussion. While Handoff shares denoiser switching with T-Stitch (Pan et al., 2024), we study how intermediate generative states support collaboration between unified multimodal models and specialized experts. Beyond generation, we show that a text-to-image expert can continue an editing trajectory without direct source-image conditioning. P-OPD further transfers expert predictions on the student's own suffix states into the unified model, enabling standalone inference.

Limitations. Our current implementation requires a shared VAE latent space and compatible noise and timestep parameterizations. Performance depends on the handoff timestep, and expert continuation may alter source details in editing. A potential future direction is to learn a mapping between intermediate states in different VAE representations, enabling collaboration across VAE.

6 Conclusion

We propose Handoff, a training-free strategy that lets a unified multimodal model and a specialized expert collaborate through an intermediate generative state. By letting the unified model establish the semantics and layout in the early sampling steps and the expert refine the rendering in the later ones, Handoff improves both image generation and editing across three unified models and four experts from two VAE families, often surpassing both standalone models. We further introduce Partial On-Policy Distillation (P-OPD), which internalizes the expert's contribution by distilling only the later part of the unified model's own trajectory, improving BAGEL on all six benchmarks without requiring the expert at inference. Importantly, our results show that intermediate generative states can serve as a lightweight interface between independently developed visual models, allowing them to evolve separately while combining their strengths.

AI USE STATEMENT

We used generative AI tools for translation and manuscript editing, and Qwen3.5-27B to synthesize training prompts. The authors reviewed the AI-assisted manuscript text and take responsibility for the final content.

ETHICS STATEMENT

This work does not involve human-subject studies or the direct collection of personal photographs. The training corpus consists of generated images, while existing public benchmarks are used only for evaluation. We will document the generation and editing provenance of the released corpus and its intended research use.

REPRODUCIBILITY STATEMENT

Section 4.1 describes the experimental settings, benchmarks, and evaluation metrics. Appendix A provides implementation details and sampling configurations for Handoff and P-OPD. Section 4.3 reports ablation studies, and Appendix B presents additional qualitative results.

REFERENCES

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2023. URL https://arxiv.org/abs/2211.01324.

Black Forest Labs. FLUX.1-dev model card. Hugging Face model repository, 2024. URL https://huggingface.co/black-forest-labs/FLUX.1-dev. Accessed 2026-08-05.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. URL https://arxiv.org/abs/2211.09800.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. OneIG-Bench: Omni-dimensional nuanced evaluation for image generation. *arXiv preprint arXiv:2506.07977*, 2025a.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation, 2025b. URL https://arxiv.org/abs/2506.07977.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters, 2023. URL https://arxiv.org/abs/2305.10855.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025a. URL https://arxiv.org/abs/2505.09568.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025b. URL https://arxiv.org/abs/2501.17811.

Shenggan Cheng, Yuanxin Wei, Lansong Diao, Yong Liu, Bujiao Chen, Lianghua Huang, Yu Liu, Wenyuan Yu, Jiangsu Du, Wei Lin, and Yang You. SRDiffusion: Accelerate video diffusion inference via sketching–rendering cooperation. *arXiv preprint arXiv:2505.19151*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. 2025. URL https://arxiv.org/abs/2505.14683.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, Yan Li, Yubo Wang, Zhijie Cao, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Yuwei Niu, Yue Zhu, Bo Liu, Chengguang Lv, Haojia Yu, Haozhe Xie, Hongli Wang, Jianan Fan, Jiaqi Li, Jiefan Lu, Jingcheng Ni, Junxiang Xu, Kaihuan Liang, Lianqiang Shi, Linjun Dai, Linyan Wang, Oscar Qian, Peng Gao, Pengfei Liu, Qingping Sun, Rui Shen, Ruisi Wang, Shengnan Ma, Shuang Yang, Siyi Xie, Siying Li, Tianbo Zhong, Xiangli Kong, Xuanke Shi, Yang Gao, Yongqiang Yao, Yves Wang, Zhengqi Bai, Zhengyu Lin, Zixin Yin, Wenxiu Sun, Ruihao Gong, Quan Wang, Lewei Lu, Lei Yang, Ziwei Liu, and Dahua Lin. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture, 2026. URL https://arxiv.org/abs/2605.12500.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Weixi Feng, Wanrong Zhu, Tsu jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models, 2023. URL https://arxiv.org/abs/2305.15393.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xuanda Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Xin Xia, Xuefeng Xiao, Zhonghua Zhai, Xinyu Zhang, Qi Zhang, Yuwei Zhang, Shijia Zhao, Jianchao Yang, and Weilin Huang. Seedream 3.0 technical report, 2025. URL https://arxiv.org/abs/2504.11346.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation, 2025. URL https://arxiv.org/abs/2404.14396.

Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, Linus, Di Wang, and Jie Jiang. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again, 2025. URL https://arxiv.org/abs/2507.22058.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022. URL https://arxiv.org/abs/2208.01626.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models, 2023. URL https://arxiv.org/abs/2305.17216.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models, 2024. URL https://arxiv.org/abs/2305.13655.

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, 2025a. URL https://arxiv.org/abs/2504.17761.

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, 2025b. URL https://arxiv.org/abs/2504.17761.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2025. URL https://arxiv.org/abs/2411.07975.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022. URL https://arxiv.org/abs/2108.01073.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022. URL https://arxiv.org/abs/2211.09794.

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

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries, 2025. URL https://arxiv.org/abs/2504.06256.

Zizheng Pan, Bohan Zhuang, De-An Huang, Weili Nie, Zhiding Yu, Chaowei Xiao, Jianfei Cai, and Anima Anandkumar. T-stitch: Accelerating sampling in pre-trained diffusion models with trajectory stitching, 2024. URL https://arxiv.org/abs/2402.14167.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Wujian Peng, Lingchen Meng, Yuxuan Cai, Xianwei Zhuang, Yuhuan Yang, Rongyao Fang, Chenfei Wu, Junyang Lin, Zuxuan Wu, and Shuai Bai. Unified multimodal autoregressive modeling with shared context-visual tokenizer is key to unification, 2026. URL https://arxiv.org/abs/2606.18249.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face, 2023. URL https://arxiv.org/abs/2303.17580.

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, 2023. URL https://arxiv.org/abs/2311.10089.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation, 2025. URL https://arxiv.org/abs/2412.15188.

Ying Tai, Nikai Du, Rui Xie, Zhennan Chen, Qian Wang, Zhengkai Jiang, Kai Zhang, and Jian Yang. Investigating text insulation and attention mechanisms for complex visual text generation, 2026. URL https://arxiv.org/abs/2503.23461.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. URL https://arxiv.org/abs/2405.09818.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, Zirun Zhu, Ziqian Fan, Leyao Gu, Haomin Wang, Qi Wei, Jinhui Yin, Xue Yang, Zhihang Zhong, Qi Qin, Yi Xin, Bin Fu, Yihao Liu, Jiaye Ge, Qipeng Guo, Gen Luo, Hongsheng Li, Yu Qiao, Kai Chen, and Hongjie Zhang. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing, 2026. URL https://arxiv.org/abs/2603.09877.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning, 2024. URL https://arxiv.org/abs/2412.14164.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing, 2024. URL https://arxiv.org/abs/2311.03054.

Alex Jinpeng Wang, Dongxing Mao, Jiawei Zhang, Weiming Han, Zhuobai Dong, Linjie Li, Yiqi Lin, Zhengyuan Yang, Libo Qin, Fuwei Zhang, Lijuan Wang, and Min Li. Textatlas5m: A large-scale dataset for dense text image generation, 2025. URL https://arxiv.org/abs/2502.07870.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024a. URL https://arxiv.org/abs/2409.18869.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing, 2024b. URL https://arxiv.org/abs/2407.05600.

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.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation, 2024a. URL https://arxiv.org/abs/2410.13848.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. OmniGen2: Towards instruction-aligned multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Junfeng Wu, Yi Jiang, Chuofan Ma, Yuliang Liu, Hengshuang Zhao, Zehuan Yuan, Song Bai, and Xiang Bai. Liquid: Language models are scalable and unified multi-modal generators, 2025c. URL https://arxiv.org/abs/2412.04332.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm, 2024b. URL https://arxiv.org/abs/2309.05519.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. Vila-u: a unified foundation model integrating visual understanding and generation, 2025d. URL https://arxiv.org/abs/2409.04429.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers, 2024. URL https://arxiv.org/abs/2410.10629.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation, 2025a. URL https://arxiv.org/abs/2408.12528.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models, 2025b. URL https://arxiv.org/abs/2506.15564.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms, 2024. URL https://arxiv.org/abs/2401.11708.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark, 2025. URL https://arxiv.org/abs/2505.20275.

Yanbing Zeng, Jia Wang, Hanghang Ma, Junqiang Wu, Jie Zhu, Xiaoming Wei, and Jie Hu. Forge-and-quench: Enhancing image generation for higher fidelity in unified multimodal models. *arXiv preprint arXiv:2601.04706*, 2026.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing, 2024a. URL https://arxiv.org/abs/2306.10012.

Xinchen Zhang, Ling Yang, Yaqi Cai, Zhaochen Yu, Kai-Ni Wang, Jiake Xie, Ye Tian, Minkai Xu, Yong Tang, Yujiu Yang, and Bin Cui. Realcompo: Balancing realism and compositionality improves text-to-image diffusion models, 2024b. URL https://arxiv.org/abs/2402.12908.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model, 2024. URL https://arxiv.org/abs/2408.11039.

A Implementation Details

A.1 Training-Free Handoff

For BAGEL and OmniGen2, both generation and editing use $K = 50$ sampling steps with a timestep shift of 3.0. We hand off at $h = 10$ for generation and $h = 30$ for editing, regardless of the editing expert destination. InternVL-U uses its default $K = 20$ sampling schedule for both tasks, with $h = 4$ for generation and $h = 9$ for editing. The handoff boundaries are fixed across test examples.

For generation, the unified model and generation expert receive the same text prompt through their respective text encoders. For editing, the unified model receives the source image and editing instruction. When handing off to an editing expert, the expert receives both the source image and instruction. When handing off to a generation expert, the expert receives only the instruction; source-image information is carried by the transferred latent. In each case, the expert continues from $(\mathbf{z}_h, t_h)$ along the remaining timesteps of the original schedule. Models within each pair share a VAE latent space, with sampling times and latent conventions aligned at handoff.

Model-specific inference configurations are summarized in Table 4.

Table 4: Training-free handoff configurations for generation and editing. The FLUX-VAE pairs use a timestep shift of 3.0; InternVL-U uses its default sampling schedule.

Pair	Task	Steps	Handoff
BAGEL → FLUX.1-dev	Gen.	50	10
OmniGen2 → FLUX.1-dev	Gen.	50	10
BAGEL → FLUX.1-Kontext	Edit	50	30
BAGEL → FLUX.1-dev	Edit	50	30
OmniGen2 → FLUX.1-Kontext	Edit	50	30
OmniGen2 → FLUX.1-dev	Edit	50	30
InternVL-U → Qwen-Image	Gen.	20	4
InternVL-U → Qwen-Image-Edit	Edit	20	9

A.2 P-OPD Training Details

For P-OPD, all pretrained parameters are frozen and only LoRA adapters are optimized. We use $K = 50$ sampling steps with suffix boundary $h = 10$ for training and $h = 30$ for editing. The LoRA rank is $r = 256$, with scaling factor $\alpha = 512$ and zero dropout. Adapters are applied to the generation-MoE attention projections.

We collect fresh on-policy rollouts at every optimization step and supervise all 40 suffix states, $\mathcal{S} = \{10, \ldots, 49\}$, using velocity targets from a frozen FLUX.1-dev expert. Training runs for 1,200 steps with AdamW, a constant learning rate of 1×10^{-4}, zero weight decay, BF16 precision, a global batch size of three trajectories, and gradient clipping at 1.0.

Training data The 1,600 generation prompts are synthesized by Qwen3.5-27B and consist of 800 text-rendering prompts (long-form and multi-region text) and 800 general prompts (compositional and knowledge-grounded scenes), without using any benchmark prompts.

A.3 State influence.

Figure 8 illustrates corresponding effects on generation and editing. For editing, a mismatched state leads to a blue goat despite the instruction requesting a yellow one. These results suggest that the state carries task-specific content that influences expert continuation alongside the textual instruction, including source-related information that enables editing without direct source-image input.

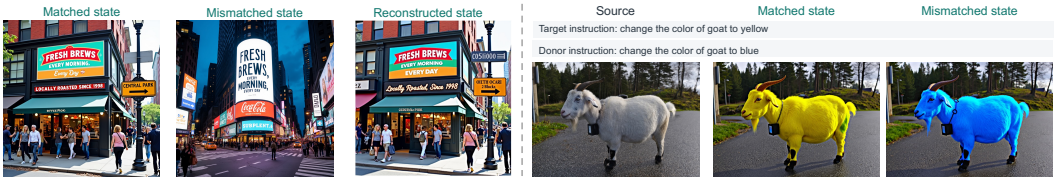

Figure 8: **Influence of the transferred state.** The receiver prompt is fixed while the transferred state is varied. Left: generation with matched, mismatched, and reconstructed states. Right: the target instruction changes the goat to yellow, while the mismatched state comes from an instruction to change it to blue.

Table 5: **Effect of the transmitted state on LongText-Bench.**

State variant	Target Recall ↑	Donor-only Recall ↓
Flux	0.612	-
Matched state	0.655	0.014
Reconstructed state	0.629	0.023
Shuffled on-manifold	0.561	0.037
Norm-matched random	0.008	0.000

B Additional Qualitative Results

B.1 Generation cases.

In Fig. 9, we visualize the performance of FLUX, BAGEL, InternVL-U, Qwen-Image, BAGEL-Handoff, OmniGen2-Handoff, and P-OPD on LongText-Bench and OneIG-Bench. These qualitative examples reveal several notable patterns that are consistent with our quantitative findings. First, unified multimodal models (UMMs), such as BAGEL, OmniGen2, and InternVL-U, often struggle with text-intensive generation scenarios, exhibiting blurry typography, incomplete rendering, or incorrect textual content. In contrast, specialized text-to-image models such as FLUX and Qwen-Image generally produce images with higher visual fidelity. However, their relatively limited prompt-understanding capability can lead to noticeable discrepancies between the generated content and the input prompt, particularly when handling long or compositionally complex instructions.

B.2 Editing cases.

We present qualitative results for all settings on GEdit and ImgEdit-Bench in Figs.11 and 12.

B.3 Failure cases.

As shown in Fig 13., UMMs exhibit several recurring failure modes in image generation, including incorrect text rendering, erroneous spatial relationships, incorrect object attributes, and physically implausible object arrangements. Notably, these failures often occur even when the model correctly captures the high-level semantics of the prompt. For example, the requested objects may all be present, yet their attributes, relative positions, or physical interactions remain inconsistent with real-world expectations. This suggests that strong semantic understanding alone is insufficient for faithful image synthesis: UMMs may understand *what* should appear in the scene while still struggling with *how* those elements should be visually instantiated and physically organized. In contrast, specialized generation models generally exhibit stronger visual and generative priors, producing more coherent object structures, realistic details, and physically plausible compositions. These observations further motivate combining the semantic understanding of UMMs with the generation capability of specialized experts.

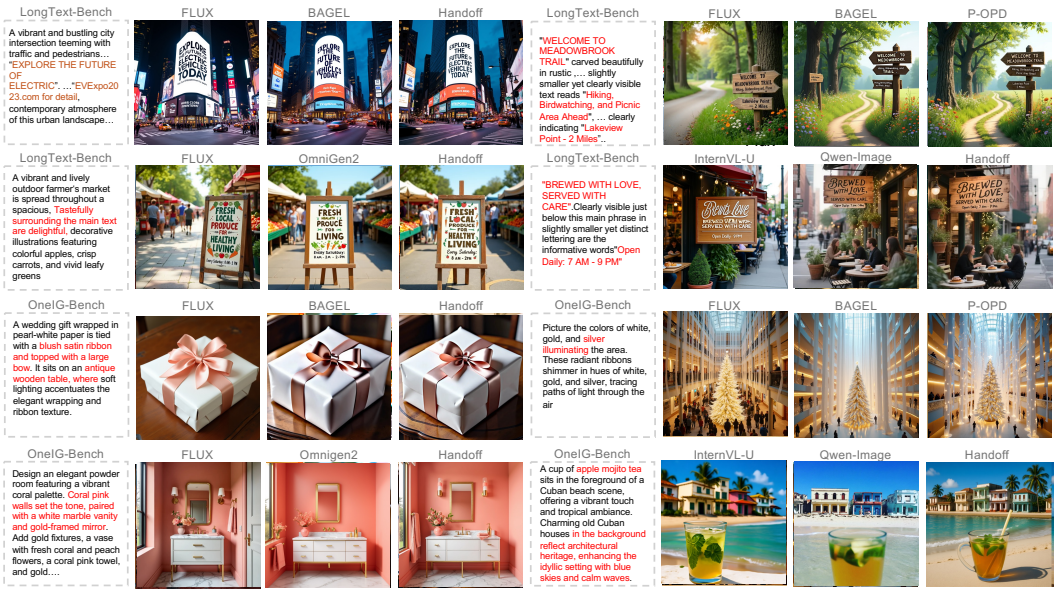

Figure 9: **Qualitative results on LongText-Bench and OneIG-Bench.** Benchmark names indicate the source of each prompt, while text highlighted in red denotes key information in the prompt or text that is expected to be rendered in the generated image.

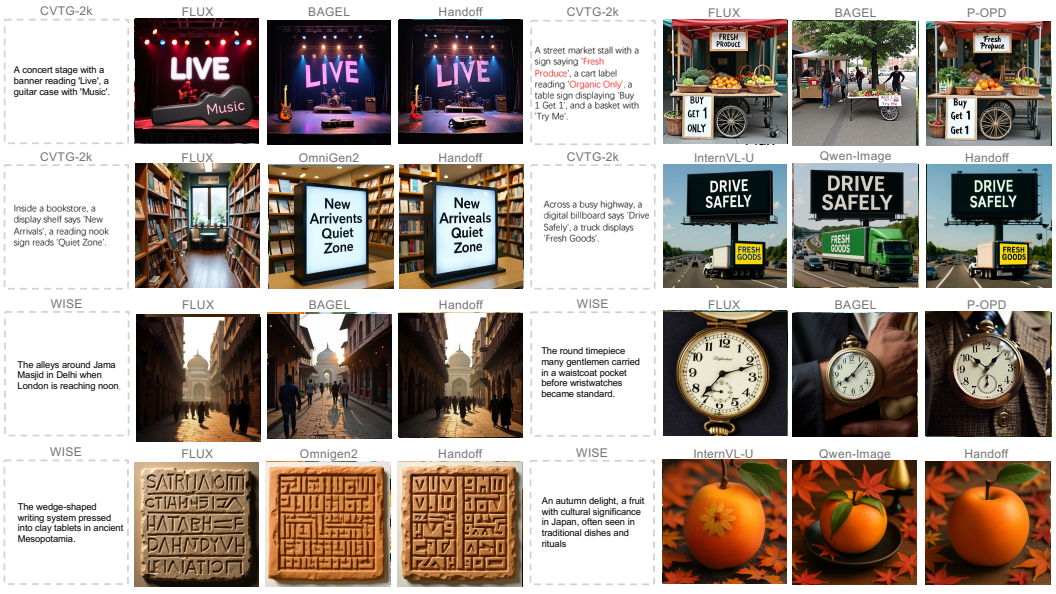

Figure 10: **Qualitative results on CVTG-2K and WISE.** Benchmark names indicate the source of each prompt, while text highlighted in red denotes key information in the prompt or text that is expected to be rendered in the generated image.

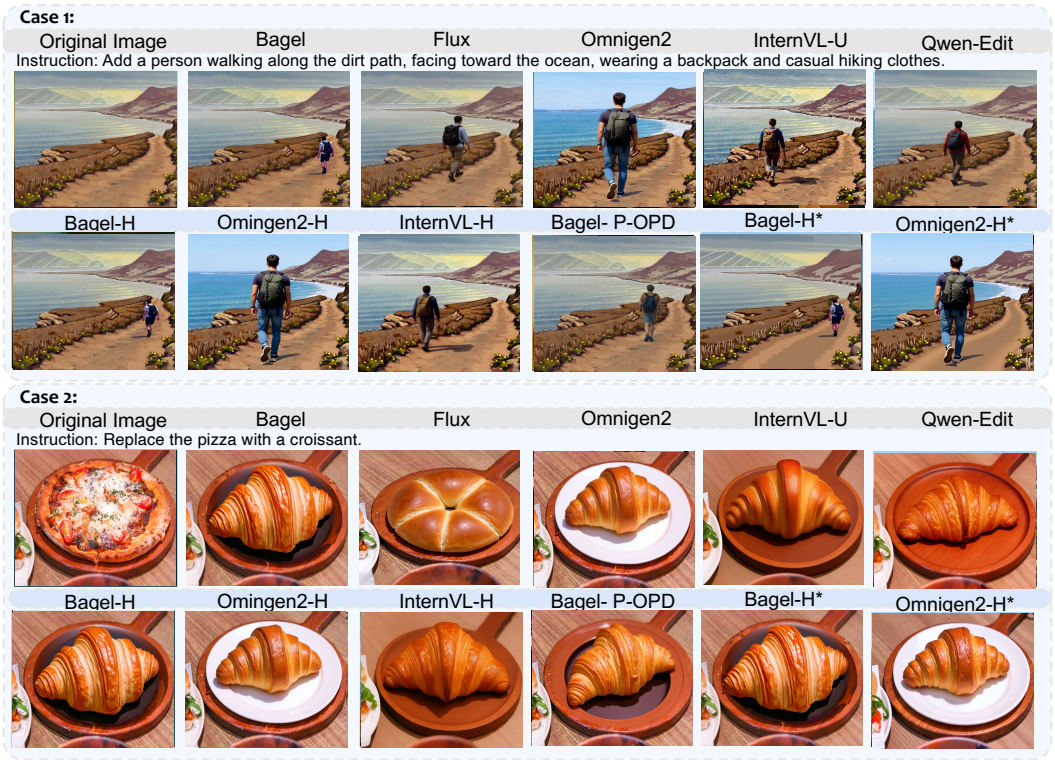

Figure 11: **Qualitative results on GEdit and ImgEdit.** The suffixes -H and -H* denote handoff to an image-editing expert and a text-to-image generation expert, respectivel.



Figure 12: **Qualitative results on GEdit and ImgEdit.** The suffixes -H and H* denote handoff to an image-editing expert and a text-to-image generation expert, respectivel.w

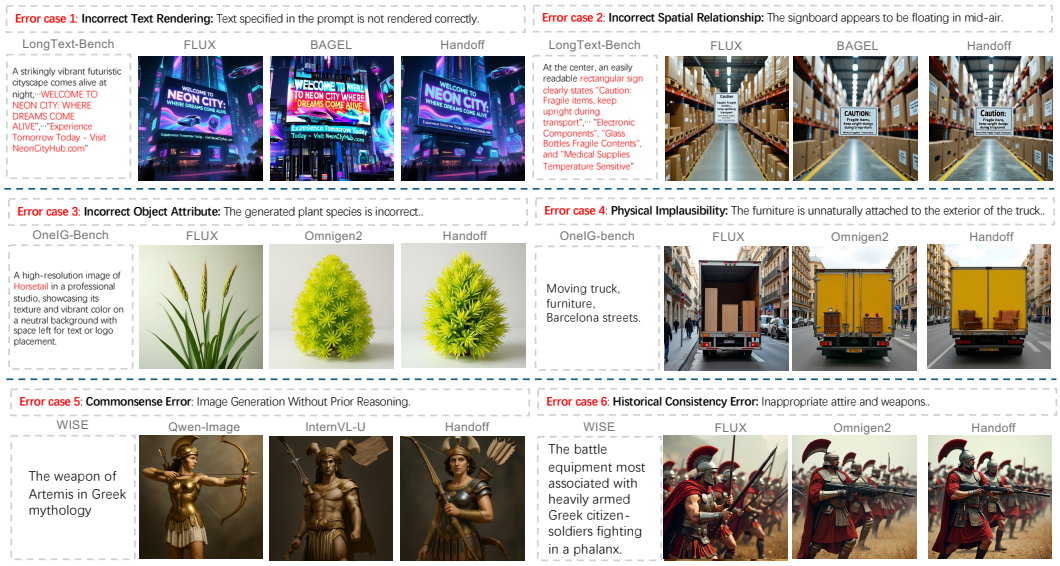

Figure 13: **Failure case of generation.** UMMs exhibit several characteristic errors, including incorrect text rendering, spatial relationship errors, incorrect object attributes, and physically implausible object arrangements. These examples reveal that strong semantic understanding does not always translate into faithful and physically realistic image synthesis.

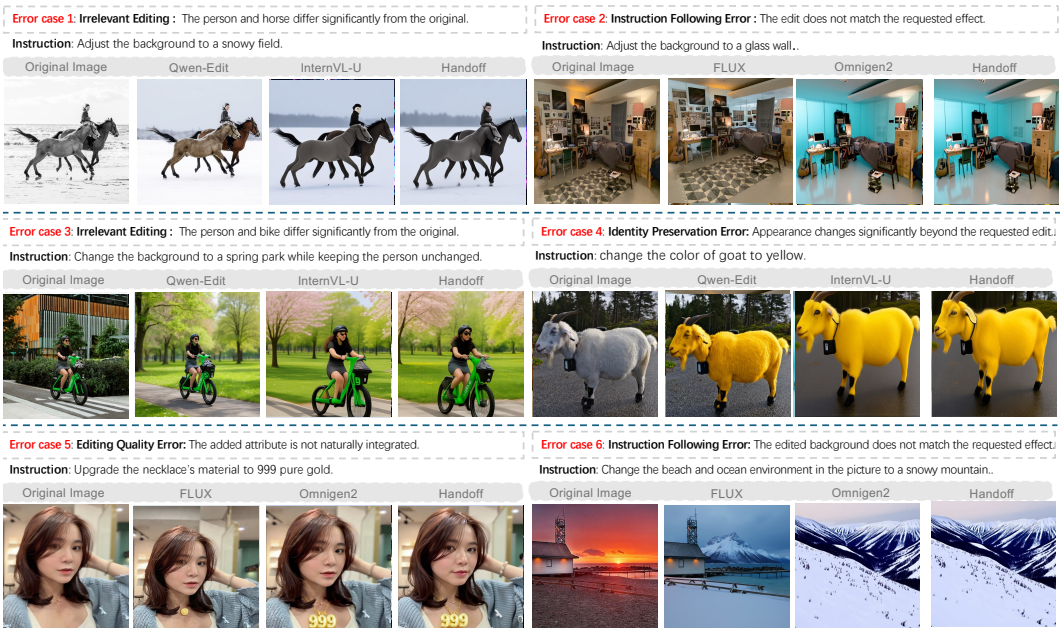

Figure 14: **Qualitative failure cases in image editing**. Common errors include imperfect instruction following, unnecessary modifications, poor identity/content preservation, and inaccurate attribute or background edits.