FUSE: Fine-Grained and Semantic-Aware Learning for Unified Image Understanding and Generation

Peng Zhang[1,2*], Wanggui He[2*†], Mushui Liu[1,2*], Wenyi Xiao[1,2], Siyu Zou[2], Yuan Li[1], Xingjian Wang[1,2], Guanghao Zhang[2], Yanpeng Liu[2], Weilong Dai[2], Jinlong Liu[2], Shuyi Ying[1], Ruikai Zhou[2], Yunlong Yu[1], Yubo Tao[1], Hai Lin[1‡], Hao Jiang[2‡]

[1]Zhejiang University, China
[2]Alibaba Group, China

12321217@zju.edu.cn, wanggui.hwg@taobao.com, {lms,wenyixiao}@zju.edu.cn, 31520211154121@stu.xmu.edu.cn, {yuanli, xingjianwang}@zju.edu.cn, guanghao.zgh@taobao.com, jingshen.lyp@alibaba-inc.com, junmu.dwl@taobao.com, LJLwykqh@126.com, sy3977@nyu.edu, 1323268276@sjtu.edu.cn, yuyunlong@zju.edu.cn, taoyubo@cad.zju.edu.cn, lin@cad.zju.edu.cn,aoshu.jh@alibaba-inc.com

Abstract

Recent unified models have demonstrated that the reasoning capacity of Multimodal Large Language Models (MLLMs) can be leveraged to facilitate diffusion-based image generation with impressive flexibility and performance. However, approaches that rely heavily on MLLMs for high-level semantic encoding often struggle with fine-grained visual tasks like image editing and virtual try-on. To address this gap, we propose FUSE, a unified framework excelling at both high-level vision–language understanding and fine-grained generation. First, we introduce a Semantic-to-Detail Connector that pre-aligns fine-grained visual features with the MLLM's semantic space. This design counteracts the low-level information loss inherent in MLLM encodings, creating a unified representation that steers the diffusion process with both global semantics and rich local details. Second, to further enhance semantic awareness and detail preservation, we introduce Adaptive-GRPO, a post-training objective that dynamically balances semantic coherence against pixel-level fidelity. The integration of these two innovations allows FUSE to generate images that are both semantically faithful and visually fine-grained. Comprehensive experiments on text-to-image and instruction-guided editing benchmarks show that FUSE significantly outperforms existing unified baselines, achieving 0.89 on Geneval, 0.65 on WISE, and 3.88 on ImageEdit.

Introduction

Recent advanced Multimodal Large Language Models (MLLMs) (Bai et al. 2025a; Zhu et al. 2025; Hurst et al. 2024) and diffusion generative models (Esser et al. 2024; Labs 2024a; Liu et al. 2025b; Ma et al. 2025a; Liu et al. 2025c; Ma et al. 2024a, 2025b) have demonstrated remarkable capabilities in both understanding and generation tasks. Despite these successes, researchers continue to pursue a unified framework that effectively integrates these two capabilities. Currently, unified frameworks explored in the community broadly fall into two approaches. The first approach involves the end-to-end integration of LLMs and diffusion models, exemplified by methods such as Transfusion (Zhou, Yu et al. 2025), JanusFlow (Ma et al. 2024b), and BAGLE (Deng et al. 2025b). Although merging understanding and generation into a unified transformer backbone promotes rich cross-modal interactions, it demands an exceptionally large, cross-modal training dataset to achieve stable alignment between modalities. Even with substantial training resources, the generated image quality from these methods typically remains inferior compared to specialized diffusion-based models like Flux (Labs 2024a).

The second approach uses MLLMs as semantic guidance combined with existing diffusion generation models. This method offers several advantages: it requires minimal training overhead while successfully injecting MLLM's understanding capabilities into existing diffusion models, enabling unified understanding and generation with high image quality. Representative methods such as MetaQuery (Pan et al. 2025) and Ming-Lite (AI et al. 2025) introduce learnable queries into the MLLM to compress high-level semantic guidance. However, this approach suffers from a critical drawback of poor detail preservation capability. The MLLM's semantic bottleneck inevitably discards low-level visual cues that are indispensable for photorealistic reconstruction. Subsequent efforts like Step1X-Edit (Liu et al. 2025d) attempt to address this by directly feeding raw VAE latents to the diffusion branch, but the network often fails to properly fuse these modalities, falling into shortcut solutions that simply copy input latents rather than learning reliable editing operators. *These limitations highlight a fundamental challenge: How can unified models maintain high-quality image generation while simultaneously preserving fine-grained details?*

To address this challenge, we introduce **FUSE**, a unified framework that achieves semantic-aware, high-quality image generation with excellent detail preservation through

*These authors contributed equally.
†Project lead.
‡Corresponding author.

Figure 1: FUSE exhibits remarkable semantic coherence and fidelity in image generation, while also showcasing strong instruction-following proficiency in image editing and virtual try-on.

innovative model architecture and loss functions. First, we devise a **Semantic-to-Detail Connector** (SD-Connector) that pre-fuses semantic representations from a pre-trained MLLM with low-level VAE features, yielding a single, unified representation. These entangled features inherit high-level semantics while retaining rich low-level visual cues, yet avoid exposing the diffusion model to raw VAE features. By doing so, the SD-Connector prevents the diffusion model from over-relying on low-level shortcuts while still providing strong detail-preserving capability. To further balance the integration of semantic and VAE information and make VAE utilization task-adaptive, we embed a dual-expert transformer within the connector, granting the framework the flexibility demanded by diverse image-generation tasks.

Secondly, to further encourage the semantic–generative feature coupling required for controllable synthesis, we introduce **Adaptive-GRPO**, a post-training objective that adaptively reweights two complementary rewards: semantic coherence and low-level fidelity, at each denoising step. Unlike pure diffusion-loss objectives, which struggle to enforce subtle pixel-level changes in image-editing tasks (Wang et al. 2025a), Adaptive-GRPO exploits the intrinsic step-wise structure of the diffusion process: early timesteps encode coarse, high-level semantics, whereas later ones govern low-level visual fidelity. By aligning the reward schedule with this natural progression, our method emphasizes semantic fidelity when it is most influential (early steps) and pixel-level accuracy when it becomes dominant (late steps). The resulting unified objective seamlessly integrates semantic-level and fine-grained goals, enabling precise semantic guidance in the VAE latent space and producing outputs that align more closely with human preferences.

Overall, our contributions are summarized as follows:

1. We present **FUSE**, a unified framework that seamlessly couples high-level semantic understanding with fine-grained visual detail, enabling photorealistic generation, precise image editing, and robust visual reasoning within a single model.
2. We design a **Semantic-to-Detail Connector** which pre-fuses the high-level semantic representations extracted from the MLLM with fine-grained visual features to steer the diffusion denoising process.
3. We propose **Adaptive-GRPO**, a training strategy that offers more nuanced guidance during generation, enabling the model to reconcile semantic requirements with fine-grained fidelity within a unified objective.
4. We conduct extensive experiments demonstrating that FUSE achieves state-of-the-art performance across a diverse range of tasks, including text-to-image generation and instruction-guided editing.

Related Work

Unified Models with Diffusion Decoder. Recent unified frameworks increasingly employ diffusion decoders for image synthesis due to their strong cross-modal generation quality. Existing methods primarily fall into two categories based on how semantic information is integrated into the diffusion process. **(1) End-to-End Fusion Approaches** (Zhou,

Yu et al. 2025; Ma et al. 2024b; Deng et al. 2025b; Xie et al. 2024; Shi et al. 2024; He et al. 2025; Wang et al. 2025b) integrate semantic understanding and pixel-level generation within a unified transformer-based architecture trained jointly from scratch. These models typically encode images with auto-encoders optimized explicitly for pixel reconstruction, providing detailed visual features directly to the diffusion decoder. While this strategy yields high-fidelity imagery, the pixel-centric representation often lacks precise semantic alignment with textual prompts, limiting the model's capability to capture nuanced semantic intent. Moreover, training these tightly integrated architectures demands large-scale cross-modal datasets to stabilize modality alignment, and even then, the resulting image quality tends to lag behind specialized diffusion generators (Labs 2024a; Li, Kamko et al. 2024). **(2) Semantic-Guided Generation Approaches** (Ge et al. 2024; Lin et al. 2025; AI et al. 2025; Liu et al. 2025d) leverage pre-trained MLLMs to provide semantic conditioning for separate diffusion-based generation models. These frameworks introduce compact, semantically aligned representations—such as learnable queries (Pan et al. 2025) or vision tokens (Chen et al. 2025a) into the diffusion decoding process. While effectively injecting strong semantic coherence, these abstract representations inherently discard essential low-level visual details, impairing fine-grained detail preservation. Attempts to preserve low-level details, like Step1X-Edit(Liu et al. 2025d), by directly incorporating raw VAE latents into the diffusion model often result in trivial solutions where the model merely replicates input latents instead of performing meaningful, pixel-level edits. Thus, achieving a unified model that simultaneously balances fine-grained visual fidelity and robust semantic alignment remains an open challenge. Thus, achieving a unified model that simultaneously balances fine-grained visual fidelity and robust semantic alignment remains an open challenge.

Alignment for Image Generation Models. A growing line of research aligns pre-trained diffusion- or flow-based image generators with downstream objectives by means of reward modeling. Scalar-reward (Prabhudesai et al. 2023; Xu et al. 2023) baselines fine-tune back-propagates a per-sample reward loss only on the final denoising steps. However, the limited optimize steps of this gradient signal bounds the achievable improvement. Direct Preference Optimization (DPO) approachs (Dong et al. 2023; Wallace et al. 2024) reframe alignment as a supervised classification problem on human preference pairs. While training-efficient, they suffer from a train–test distribution shift that ultimately limits peak performance. PPO-style variants (Black et al. 2023; Gupta et al. 2025) yield stronger alignment by treating the denoising process as a sequential decision problem, yet they require a separate value network, which materially increases memory consumption and tuning overhead. Group Relative Policy Optimization (GRPO) (Shao et al. 2024), originally developed for LLMs, has emerged as a post-training recipe that dispenses with the value model while retaining optimization quality. Flow-GRPO (Liu et al. 2025a) recently adapted GRPO to flow-based generators by converting deterministic ODE trajectories into stochastic SDE counterparts. However, applying Flow-GRPO to unified generation-and-editing architectures remains challenging because every task (e.g., text-to-image generation, instruction-guided editing) demands a task-specific reward model that is carefully designed and selectively invoked (Wang et al. 2025a), rendering global alignment elusive.

Methodology

FUSE Architecture

Our proposed FUSE, the MLLM utilized is Qwen2.5-VL (Bai et al. 2025b), while the image diffusion model employed is Flux-1.0-DEV (Labs 2024a). Notably, the MLLM remains in a frozen state throughout the process. The overall framwork is illustrated in Fig. 2.

Generation Queries. MLLM encodes the textual prompt and images into textual tokens T_{token} and visual tokens V_{token}. Motived by MetaQuery (Pan et al. 2025), we also employ randomly initialized learnable queries $G \in \mathbb{R}^{N\times D}$, subsequently concatenated with the textual and visual tokens, forming $\{T_{\text{token}}, V_{\text{token}}, G\}$. These concatenated tokens are then processed through the Transformer architecture of Qwen2.5-VL, resulting in an output that integrates semantic information from the text prompt along with its original token information:

$$\{T'_{\text{token}}, V'_{\text{token}}, G'\} = \text{MLLM}(\{T_{\text{token}}, V_{\text{token}}, G\}) \quad (1)$$

where $\{T'_{\text{token}}, V'_{\text{token}}\}$ are utilized for understanding, while G' serve as the condition for generation.

Semantic-to-Detail Connector. While generation queries establish the linkage between the MLLM and the generative model, they inherently lack the capacity to preserve image details due to their high-level visual token characteristics. To address this issue, we propose the Semantic-to-Detail Connector (SD-Connector) to seamlessly integrate visual detail features (Labs 2024a), aligning them with the MLLM's semantic features G' as defined in Eq. 1. Specifically, we employ the VAE (Labs 2024a) to extract low-level image features F_V, which are subsequently fused with the generation queries G'. The SD-Connector then utilizes a dual-stream transformer architecture to process F_V and G' separately:

$$\begin{aligned}
Q_{F_V}, K_{F_V}, V_{F_V} &= W_Q^V(F_V), W_K^V(F_V), W_V^V(F_V) \\
Q_{G'}, K_{G'}, V_{G'} &= W_Q^{G'}(F_{G'}), W_K^{G'}(F_{G'}), W_V^{G'}(F_{G'}) \\
\{\widetilde{F_V}, \widetilde{F_{G'}}\} &= \text{ATTN}\big(\{Q_{F_V}, Q_{G'}\}, \{K_{F_V}, K_{G'}\}, \{V_{F_V}, V_{G'}\}\big) \\
\widehat{F_V} &= \text{FFN}_V(\widetilde{F_V}), \quad \widehat{F_{G'}} = \text{FFN}_{G'}(\widetilde{F_{G'}})
\end{aligned} \quad (2)$$

where W_Q, W_K, W_V are learnable linear projection matrices, ATTN denotes the self-attention mechanism, $\{\}$ denotes the concat operation, and FFN is a Feed-Forward Network, where the scripts V and G' denote the parameters for each respective stream.

The resulting unified token sequence $\{\widehat{F}_V; \widehat{F}_{G'}\}$ combines global semantics with rich local details and serves as the unifed conditioning input for the diffusion decoder.

Generation Model. We plug this unified feature sequence into the FLUX-DEV diffusion backbone, replacing its original text embedding conditioned for the denosing process.

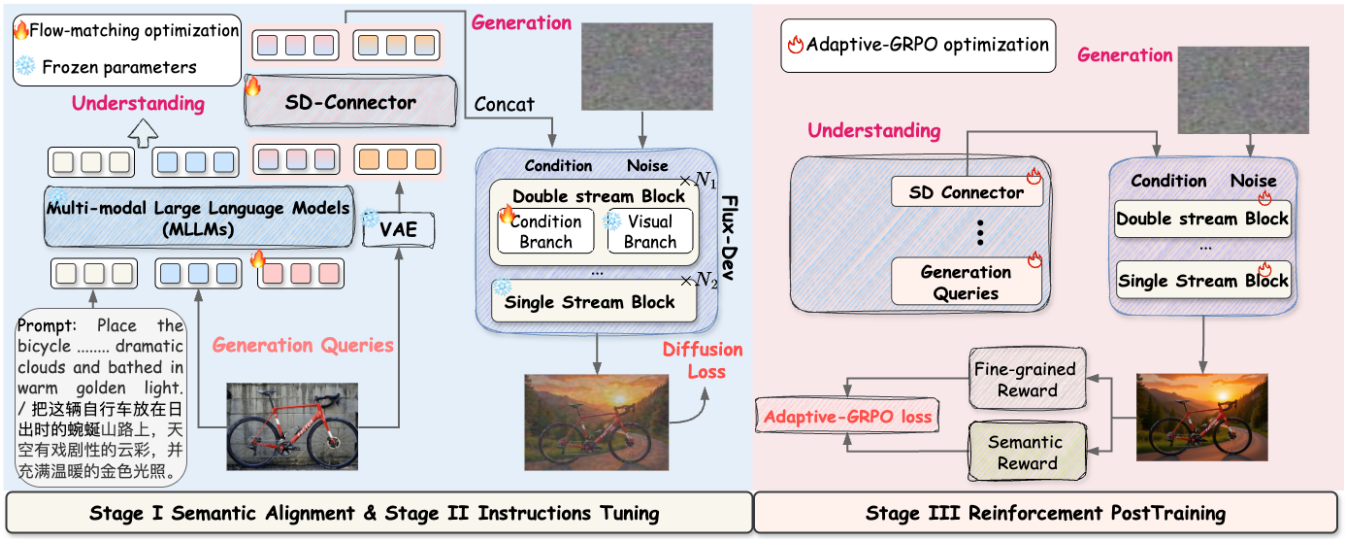


Figure 2: Architecture and Training pipeline of FUSE. We introduce an SD-Connector that unifies semantic-level and fine-grained visual features to steer the diffusion process. A three-stage training curriculum progressively strengthens the model's semantic understanding and fine-grained generation capability.

The image generation of FUSE is defined as:

$$\mathbf{I}_{\text{out}} = \text{FLUX-DEV}\big(z, \{\widehat{F}_V; \widehat{F}_{G'}\}\big), z \sim \mathcal{N}(0, \mathbf{I}), \quad (3)$$

where $\mathbf{I}_{\text{out}}$ is the synthesized image conditioned on $\{\widehat{F}_V; \widehat{F}_{G'}\}$. and z is a noise sampled from the standard normal distribution.

Adaptive GRPO

Despite our model architecture being designed to simultaneously accommodate both semantic and fine-grained visual generation information, we observe that supervising the image generation process solely through flow-matching loss remains insufficient for the model to effectively leverage these two distinct modalities of information. This limitation is particularly evident in tasks requiring localized pixel-level modifications guided by semantic guidance. To resolve this limitation, we propose Adaptive-GRPO, a training objective that adapts the Flow-GRPO (Liu et al. 2025a) algorithm specifically for unified semantic and fine-grained image generation tasks.

We first train Generative Reward Models (GRMs) using the strong semantic priors of large vision-language models, thereby ensuring accurate evaluation of generation quality across different contexts, including text-to-image and image-editing scenarios. Unlike traditional scalar-only reward modeling, our GRM employs chain-of-thought reasoning to produce nuanced, multidimensional assessments. Specifically, our GRM outputs two distinct yet complementary normalized scores: a semantic coherence score $r^{\text{sem}}(x_0, c) \in [0, 1]$, and an fine-grained score $r^{\text{qual}}(x_0, c) \in [0, 1]$. Formally, let ϕ denote GRM parameters and $\mathcal{D}$ represent the dataset comprising image–prompt pairs alongside human-generated critiques. We optimize the GRM by minimizing the joint negative log-likelihood loss:

$$\mathcal{L}_{\text{GRM}} = -\mathbb{E}_{\mathcal{D}}\Big[\log p_\phi\big(y^{\text{sem}}, y^{\text{qual}}, r^{\text{sem}}, r^{\text{qual}} \mid x_0, c\big)\Big], \quad (4)$$

where $(y^{\text{sem}}, y^{\text{qual}})$ represent chain-of-thought textual justifications provided by human annotators, and $(r^{\text{sem}}, r^{\text{qual}})$ are their corresponding scalar reward labels. At inference, only these scalar semantic and perceptual rewards guide Adaptive-GRPO.

Secondly, utilizing these fine-grained reward signals, we devise an **step-wise credit assignment** strategy. This design exploits a fundamental property intrinsic to diffusion processes: earlier denoising steps primarily determine the semantic structure of the generated image, whereas later stages are crucial for refining fine-grained visual details. To capture this progressive refinement explicitly, We redistribute the rewards across all timesteps $t = 0, \ldots, T$ using time-dependent weights:

$$w^{\text{sem}}(t) = \exp\big(-\alpha\, t/T\big), w^{\text{qual}}(t) = 1 - w^{\text{sem}}(t), \quad (5)$$

where $\alpha > 0$ is a tunable hyperparameter that controls how strongly semantics dominate the early phase. Consequently, each timestep receives a composite reward:

$$R_t(x_t, a_t, c) = w^{\text{sem}}(t)\; r^{\text{sem}}(x_0, c) + w^{\text{qual}}(t)\; r^{\text{qual}}(x_0, c). \quad (6)$$

By integrating this timestep-aware reward assignment, Adaptive-GRPO converts initially coarse global rewards into fine-grained, step-wise feedback signals, effectively aligning semantic coherence and pixel-level accuracy throughout the diffusion process. The final generative policy is optimized by maximizing the following objective:

$$J_\theta = \mathbb{E}_{c,\{x^i\}}\Big[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\min\Big(r_t^i(\theta)\hat{A}_t^i, \text{clip}\big(r_t^i(\theta), 1-\varepsilon, 1+\varepsilon\big)\hat{A}_t^i\Big)\Big]. \quad (7)$$

Model	Model Type	Single Obj.↑	Two Obj.↑	Counting↑	Colors↑	Position↑	Color Attr.↑	Overall↑
PixArt-α	Gen. Only	0.98	0.50	0.44	0.80	0.08	0.07	0.48
SDXL	Gen. Only	0.98	0.74	0.39	0.85	0.15	0.23	0.55
DALL-E 3	Gen. Only	0.96	0.87	0.47	0.83	0.43	0.45	0.67
SD3-Medium	Gen. Only	**0.99**	0.94	0.72	0.89	0.33	0.60	0.74
FLUX-1-dev	Gen. Only	0.98	0.93	0.75	0.93	0.68	0.65	0.82
Janus	Unified	0.97	0.68	0.30	0.84	0.46	0.42	0.61
Emu3-Gen	Unified	0.99	0.81	0.42	0.80	0.49	0.45	0.66
Show-o	Unified	0.98	0.80	0.66	0.84	0.31	0.50	0.68
Janus-Pro-7B	Unified	**0.99**	0.89	0.59	0.90	**0.79**	0.66	0.80
MetaQuery-XL	Unified	–	–	–	–	–	–	0.80
BLIP3-o	Unified	–	–	–	–	–	–	0.84
BAGEL	Unified	0.98	0.95	0.84	**0.95**	0.78	**0.77**	0.88
UniWorld-V1	Unified	**0.99**	0.93	0.79	0.89	0.49	0.70	0.80
Ours	Unified	**0.99**	**0.98**	**0.88**	**0.96**	0.77	0.76	**0.89**

Table 1: Comparison with different models on the GenEval Benchmark. "Gen. Only" indicates generation-only models, whereas "Unified" denotes models equipped with both understanding and generation capabilities.

$r_t^i(\theta) = \dfrac{p_\theta(x_{t-1}^i \mid x_t^i, c)}{p_{\theta_{\text{old}}}(x_{t-1}^i \mid x_t^i, c)}$ and the advantage $\hat{A}_t^i$ is computed from the per-step reward $R_t(x_t^i, a_t^i, c)$ defined above.

Training Strategy

Stage I Semantic Alignment. In this initial stage, we train the T2I task on low-resolution images to align semantic representations between the MLLM and the diffusion model. Only a subset of high-quality open-source text-to-image datasets is used. All images are resized to 256^2 pixels. To avoid degrading generation quality, only the SD-Connector and text blocks of FLUX are optimized, with all other parameters kept frozen. The flow-matching objective (Liu, Gong, and Liu 2022) is used as the training objective.

Stage II Instruction Tuning. With semantic alignment established, we jointly train text-to-image (T2I) generation and image editing tasks using a curated mixture of high-quality T2I and editing datasets. We adopt a task mixing ratio of 3:7 (T2I : Edit). For both tasks, we filter out images whose shorter side is below 512 pixels, and cap the maximum resolution at 1024^2. The optimization modules and training objectives remain unchanged from Stage I.

Stage III Adaptive-GRPO Post-training. After the two-stage training, the model demonstrates strong capabilities in both text-to-image generation and image editing. Nonetheless, it may still face slight limitations in performing fine-grained pixel-level edits. To further enhance the model's performance, we utilize the proposed Adaptive-GRPO to improve both generation fidelity and controllability. We employ the trained GRM to jointly optimize semantic and perceptual rewards across both tasks using a 1:1 task ratio, while fine-tuning the SD-Connector and the full Flux backbone under the Adaptive-GRPO objective defined in Eq. (7).

Model	**WISE↑**	**MJHQ-30K FID↓**
Pure Generation Models		
PixArt-α	0.47	7.9
Emu3-Gen	–	7.8
SDXL	0.43	7.4
DALL-E 3	–	7.1
SD3-Medium	0.46	6.8
FLUX.1-dev	0.50	6.6
playground-v2.5	0.49	6.9
Unified Models		
Janus	0.23	13.2
Emu3	0.39	11.0
Show-o	0.35	9.5
Janus-Pro-7B	0.35	8.4
MetaQuery-XL	0.55	6.1
BLIP3-o	–	6.8
BAGEL	0.52	6.4
UniWorld-V1	0.55	6.3
Ours	**0.65**	**5.6**

Table 2: Performance on WISE and MJHQ-30K.

Experiments

Evaluation of Image Generation

Benchmark. We conduct a thorough evaluation of our model using established T2I generation benchmarks, e.g., GenEval (Ghosh et al. 2023), WISE (Niu, Ning et al. 2025), MJHQ30K (Li, Kamko et al. 2024).

Quantitative and Qualitative Results. The comparison methods include the image generation models PixArt-α (Chen et al. 2023), Emu3-Gen (Wang et al. 2024), DALL-E 3 (Shi et al. 2020), SD3-Medium (Esser et al. 2024), FLUX.1-dev (Labs 2024b), Playground-v2.5 (Liu et al. 2024); and the unified models Janus (Wu et al. 2024),

Figure 3: FUSE demonstrates proficiency across various tasks, including text-to-image generation, multiple image editing tasks, and virtual try-on. It achieves high semantic adherence and pixel-level consistency.

Emu3 (Wang et al. 2024), Show-o (Xie et al. 2024), Janus-Pro-7B (Chen et al. 2025b), MetaQuery-XL (Pan et al. 2025), BLIP3-o (Chen et al. 2025a), BAGEL (Deng et al. 2025a), and UniWorld-V1 (Lin et al. 2025). FUSE demonstrates exceptional prompt-following capabilities, significantly outperforming existing models on the GenEval benchmark, achieveing a superior overall score of 0.89. Its proficiency is evident across diverse prompt categories, achieving near-perfect scores of 0.99 for Single Object and 0.98 for Two Object prompts. Furthermore, FUSE excels in more challenging compositional tasks, scoring 0.88 in Counting and 0.96 in Color, consistently surpassing contemporary models. In addition to prompt adherence, FUSE produces images of outstanding visual quality. On the MJHQ-30K benchmark, the model inherits the powerful generation foundation of Flux and further enhances it through fine-tuning with the adaptive-GRPO objective, achieving an impressive FID score of 5.6. This demonstrates its formidable generation performance. Concurrently, the model showcases strong world-knowledge reasoning inherited from its underlying MLLM, attaining a WISE score of 0.65. We attribute these improvements to the well-designed connector module and adaptive-GRPO optimization.

Evaluation on Image Editing

We evaluate our proposed model in ImgEditing (Ye, He et al. 2025) benchmark. The compare methods including the MagicBrush (Zhang et al. 2023), Instruct-P2P (Brooks et al. 2023), AnyEdit (Yu et al. 2024), UltraEdit (Zhao et al. 2024), Step1X-Edit (Liu et al. 2025e), BAGEL (Deng et al. 2025a), UniWorld-V1 (Lin et al. 2025). For the image editing task, we evaluated FUSE in terms of both prompt-following capability and pixel-level consistency, which are essential for various downstream applications. The results on the ImageEdit benchmark are summarized in Table 3. FUSE outperforms existing models, achieving an impressive overall score of 3.88. Specifically, it demonstrates exceptional performance across multiple editing categories, attaining scores of 4.11, 4.41, 3.98, and 4.66 in the Add, Replace, Remove, and Action tasks, respectively. Notably, FUSE exhibits superior proficiency in handling complex image manipulations, thus surpassing many contemporary ap-

Model	Add↑	Adjust↑	Extract↑	Replace↑	Remove↑	Style↑	Action↑	Hybrid↑	Bkgd.↑	Overall↑
MagicBrush	2.72	1.47	1.31	1.89	1.57	2.49	1.39	1.80	2.03	1.85
Instruct-P2P	2.29	1.79	1.33	1.93	1.49	3.54	1.51	1.48	1.67	1.89
AnyEdit	3.12	2.66	1.82	2.71	2.34	3.27	3.31	2.07	2.37	2.63
UltraEdit	3.63	3.01	2.02	3.13	1.71	3.69	3.57	2.33	3.31	2.93
Step1X-Edit	3.90	3.13	1.87	3.45	2.61	4.44	3.43	2.52	3.19	3.17
BAGEL	3.55	3.30	1.56	3.38	2.44	4.24	4.29	2.55	3.22	3.17
UniWorld-V1	3.86	3.70	2.23	3.49	3.54	4.22	3.44	3.13	2.76	3.37
Ours	**4.11**	3.59	**2.94**	**4.41**	**3.98**	4.20	**4.66**	**3.48**	**4.63**	**3.88**

Table 3: Comparison on ImgEdit-Bench. FUSE demonstrates competitive performance across multiple editing tasks.

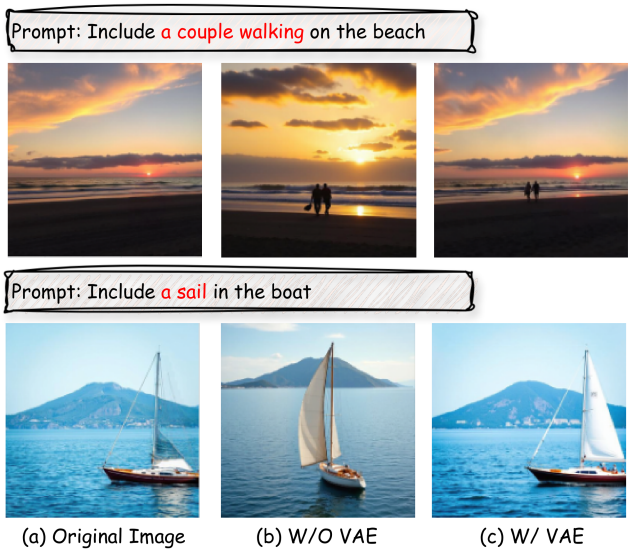
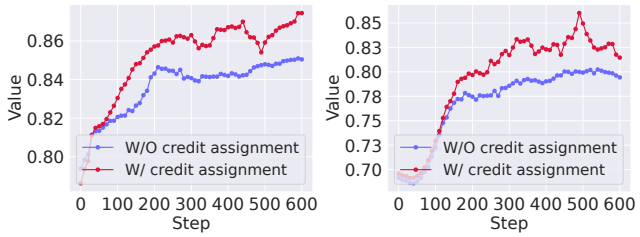


Figure 4: Effectiveness of the proposed SD-Connector. (a) Original image (b) W/O VAE. (c) W/ VAE.



(a) Semantic Reward (b) Fine-grained Reward

Figure 5: Comparison of step-wise and uniform credit assignment under the same reward type.

proaches. Furthermore, FUSE consistently preserves both pixel-level fidelity and subject identity across a wide spectrum of editing tasks—from straightforward object addition/removal and basic color or style adjustments to the more demanding scene-level style transformations. As shown in Fig. 3, the model achieves near-perfect alignment in addition operations and maintains the protagonist's appearance under drastic stylistic shifts. In virtual try-on scenarios, FUSE likewise upholds high subject consistency while rendering garments with photorealistic detail. These results showcase FUSE's superior capability in precise and faithful editing.

Ablation Studies

Feature Alignment in SD-Connector. Thanks to the flexible design of our SD-Connector, FUSE can optionally inject VAE features to incorporate low-level visual cues during generation. To assess its effectiveness in aligning semantic and VAE-derived features, we conduct ablations by toggling VAE feature injection at the diffusion stage. As shown in Fig. 4, FUSE achieves strong zero-shot editing on the any2any-add subset. Incorporating VAE features significantly improves editing consistency, while the model still performs semantic-level editing without them. This flexibility highlights the connector's ability to fuse multimodal features and adapt to varying editing demands.

Effectiveness of Step-wise Credit Assignment. To enhance semantic alignment and fine-grained generation, we introduce step-wise credit assignment into adaptive-GRPO. As shown in Fig. 5, applying step-wise reward reweighting—compared to uniform weighting across time steps—yields higher reward scores for both reward types on the test set. This demonstrates that aligning reward signals with the denoising process improves both semantic fidelity and fine-grained visual detail.

Conclusion

In this paper, we introduce FUSE, a unified MLLM–diffusion framework capable of semantic-level reasoning and fine-grained visual generation. We propose a Semantic-to-Detail Connector that integrates MLLM embeddings with VAE latents, steering the diffusion backbone with global semantics and fine-grained information. We also propose an Adaptive-GRPO objective that reallocates rewards across denoising timesteps to emphasize semantic coherence early and pixel-level fidelity late, yielding a harmonized training signal. Evaluated on Geneval, WISE, and ImageEdit, FUSE surpasses prior unified models, showing that low-level guidance can be added to MLLM-based generation without harming high-level understanding, positioning FUSE as a strong unified model for tasks requiring semantic awareness and fine-grained detail.

Acknowledgments

This research was supported in part by Zhejiang Provincial Natural Science Foundation of China under Grant LD24F020016, the Key R&D Program of Zhejiang Province 2025C01075, 2023C01043, the National Natural Science Foundation of China under Grant 62576313, 2025Z128, and Alibaba Group through Alibaba Research Intern Program.

References

AI, I.; Gong, B.; Zou, C.; Zheng, D.; Yu, H.; Chen, J.; Sun, J.; Zhao, J.; Zhou, J.; Ji, K.; Ru, L.; Wang, L.; Guo, Q.; Liu, R.; Chai, W.; Xiao, X.; and Huang, Z. 2025. Ming-Lite-Uni: Advancements in Unified Architecture for Natural Multimodal Interaction. arXiv:2505.02471.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025a. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Black, K.; Janner, M.; Du, Y.; Kostrikov, I.; and Levine, S. 2023. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*.

Brooks, T.; Holynski, A.; Efros, A. A.; et al. 2023. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 18392–18402.

Chen, J.; Xu, Z.; Pan, X.; Hu, Y.; Qin, C.; Goldstein, T.; Huang, L.; Zhou, T.; Xie, S.; Savarese, S.; et al. 2025a. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2023. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*.

Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025b. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Deng, C.; Zhu, D.; Li, K.; Gou, C.; Li, F.; Wang, Z.; Zhong, S.; Yu, W.; Nie, X.; Song, Z.; et al. 2025a. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*.

Deng, C.; Zhu, D.; Li, K.; Gou, C.; Li, F.; Wang, Z.; et al. 2025b. Emerging Properties in Unified Multimodal Pretraining. *arXiv preprint arXiv:2505.14683*.

Dong, H.; Xiong, W.; Goyal, D.; Zhang, Y.; Chow, W.; Pan, R.; Diao, S.; Zhang, J.; Shum, K.; and Zhang, T. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

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

Ge, Y.; Zhao, S.; Zhu, J.; Ge, Y.; Yi, K.; Song, L.; Li, C.; Ding, X.; and Shan, Y. 2024. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*.

Ghosh, D.; et al. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeuralIPS*, 36: 52132–52152.

Gupta, S.; Ahuja, C.; Lin, T.-Y.; Roy, S. D.; Oosterhuis, H.; de Rijke, M.; and Shukla, S. N. 2025. A simple and effective reinforcement learning method for text-to-image diffusion fine-tuning. *arXiv preprint arXiv:2503.00897*.

He, W.; Fu, S.; Liu, M.; Wang, X.; Xiao, W.; Shu, F.; Wang, Y.; Zhang, L.; Yu, Z.; Li, H.; et al. 2025. Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis. In *AAAI*, volume 39, 17123–17131.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Labs, B. F. 2024a. FLUX. https://github.com/black-forest-labs/flux.

Labs, B. F. 2024b. FLUX. https://github.com/black-forest-labs/flux.

Li, D.; Kamko, A.; et al. 2024. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*.

Lin, B.; Li, Z.; Cheng, X.; Niu, Y.; Ye, Y.; He, X.; Yuan, S.; Yu, W.; Wang, S.; Ge, Y.; Pang, Y.; and Yuan, L. 2025. UniWorld-V1: High-Resolution Semantic Encoders for Unified Visual Understanding and Generation. arXiv:2506.03147.

Liu, B.; Akhgari, E.; Visheratin, A.; Kamko, A.; Xu, L.; Shrirao, S.; Lambert, C.; Souza, J.; Doshi, S.; and Li, D. 2024. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*.

Liu, J.; Liu, G.; Liang, J.; Li, Y.; Liu, J.; Wang, X.; Wan, P.; Zhang, D.; and Ouyang, W. 2025a. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*.

Liu, M.; Ma, Y.; Yang, Z.; Dan, J.; Yu, Y.; Zhao, Z.; Hu, Z.; Liu, B.; and Fan, C. 2025b. Llm4gen: Leveraging semantic representation of llms for text-to-image generation. In *AAAI*, volume 39, 5523–5531.

Liu, M.; She, D.; Pang, J.; Huang, Q.; Ying, J.; He, W.; Hou, Y.; and Fu, S. 2025c. TFCustom: Customized Image Generation with Time-Aware Frequency Feature Guidance. In *CVPR*, 2714–2723.

Liu, S.; Han, Y.; Xing, P.; Yin, F.; Wang, R.; Cheng, W.; Liao, J.; Wang, Y.; Fu, H.; Han, C.; et al. 2025d. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*.

Liu, S.; Han, Y.; Xing, P.; Yin, F.; Wang, R.; Cheng, W.; Liao, J.; Wang, Y.; Fu, H.; Han, C.; et al. 2025e. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*.

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

Ma, Y.; Feng, K.; Zhang, X.; Liu, H.; Zhang, D. J.; Xing, J.; Zhang, Y.; Yang, A.; Wang, Z.; and Chen, Q. 2025a. Follow-Your-Creation: Empowering 4D Creation through Video Inpainting. *arXiv preprint arXiv:2506.04590*.

Ma, Y.; Liu, H.; Wang, H.; Pan, H.; He, Y.; Yuan, J.; Zeng, A.; Cai, C.; Shum, H.-Y.; Liu, W.; et al. 2024a. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia*, 1–12.

Ma, Y.; Liu, X.; Chen, X.; Liu, W.; Wu, C.; Wu, Z.; Pan, Z.; Xie, Z.; Zhang, H.; Zhao, L.; et al. 2024b. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*.

Ma, Y.; Yan, Z.; Liu, H.; Wang, H.; Pan, H.; He, Y.; Yuan, J.; Zeng, A.; Cai, C.; Shum, H.-Y.; et al. 2025b. Follow-your-emoji-faster: Towards efficient, fine-controllable, and expressive freestyle portrait animation. *arXiv preprint arXiv:2509.16630*.

Niu, Y.; Ning, M.; et al. 2025. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*.

Pan, X.; Shukla, S. N.; Singh, A.; Zhao, Z.; Mishra, S. K.; Wang, J.; Xu, Z.; Chen, J.; Li, K.; Juefei-Xu, F.; et al. 2025. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*.

Prabhudesai, M.; Goyal, A.; Pathak, D.; and Fragkiadaki, K. 2023. Aligning text-to-image diffusion models with reward backpropagation. *https://arxiv.org/abs/2310.03739*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shi, W.; Han, X.; Zhou, C.; Liang, W.; Lin, X. V.; Zettlemoyer, L.; and Yu, L. 2024. LMFusion: Adapting Pretrained Language Models for Multimodal Generation. *arXiv preprint arXiv:2412.15188*.

Shi, Z.; Zhou, X.; Qiu, X.; and Zhu, X. 2020. Improving image captioning with better use of captions. *arXiv preprint arXiv:2006.11807*.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2024. Diffusion model alignment using direct preference optimization. In *CVPR*, 8228–8238.

Wang, P.; Shi, Y.; Lian, X.; Zhai, Z.; Xia, X.; Xiao, X.; Huang, W.; and Yang, J. 2025a. SeedEdit 3.0: Fast and High-Quality Generative Image Editing. *arXiv preprint arXiv:2506.05083*.

Wang, X.; Zhang, X.; Luo, Z.; Sun, Q.; Cui, Y.; Wang, J.; Zhang, F.; Wang, Y.; Li, Z.; Yu, Q.; et al. 2024. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Wang, Y.; Liu, M.; He, W.; Zhang, L.; Huang, Z.; Zhang, G.; Shu, F.; Tao, Z.; She, D.; Yu, Z.; et al. 2025b. Mint: Multi-modal chain of thought in unified generative models for enhanced image generation. *arXiv preprint arXiv:2503.01298*.

Wu, C.; Chen, X.; Wu, Z.; Ma, Y.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; Ruan, C.; et al. 2024. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*.

Xie, J.; Mao, W.; Bai, Z.; Zhang, D. J.; Wang, W.; Lin, K. Q.; Gu, Y.; Chen, Z.; Yang, Z.; and Shou, M. Z. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *NeuralIPS*, 36: 15903–15935.

Ye, Y.; He, X.; et al. 2025. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*.

Yu, Q.; Chow, W.; Yue, Z.; Pan, K.; Wu, Y.; Wan, X.; Li, J.; Tang, S.; Zhang, H.; and Zhuang, Y. 2024. AnyEdit: Mastering Unified High-Quality Image Editing for Any Idea. *arXiv preprint arXiv:2411.15738*.

Zhang, K.; Mo, L.; Chen, W.; Sun, H.; and Su, Y. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. *NeuralIPS*, 36: 31428–31449.

Zhao, H.; Ma, X. S.; Chen, L.; Si, S.; Wu, R.; An, K.; Yu, P.; Zhang, M.; Li, Q.; and Chang, B. 2024. Ultraedit: Instruction-based fine-grained image editing at scale. *NeuralIPS*, 37: 3058–3093.

Zhou, C.; Yu, L.; et al. 2025. Transfusion: Predict the next token and diffuse images with one multi-modal model. *ICLR*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.