UNIFACET: ONE VISUAL SPACE, COMPLEMENTARY FACETS FOR IMAGE GENERATION AND EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified image generation and editing require a visual representation that preserves both semantic structure and fine-grained image details. Existing unified tokenizers often fine-tune pretrained vision foundation models (VFMs) under semantic and reconstruction supervision. However, the competing demands of these objectives make it difficult to improve reconstruction fidelity while preserving the pretrained semantic structure. We introduce **UniFacet**, a unified visual tokenizer that reformulates this problem as completing a pretrained semantic representation with reconstruction-relevant details. UniFacet keeps the VFM frozen as a semantic anchor and learns a lightweight residual branch to recover missing image-specific information, while regularizing its magnitude to obtain the necessary reconstruction details with minimal modification to the pretrained semantic structure. A lightweight projection compresses the resulting unified representation into a compact reconstruction facet for image decoding. To provide complementary semantic information for generative modeling, we derive a semantic facet from the same unified representation and combine it with the reconstruction facet as the generative target. The semantic projection is jointly optimized with a generative model during adaptation, while the reconstruction pathway remains fixed. We instantiate UniFacet in **UniFacet-1.5B**, a unified model for text-to-image generation and instruction-guided image editing. The model uses the unified representation for source-image conditioning and jointly predicts the reconstruction and semantic facets for image synthesis. Experiments demonstrate high-fidelity reconstruction with largely preserved pretrained semantic capabilities, together with strong text-to-image generation and instruction-guided image editing performance.

1 INTRODUCTION

Modern multimodal models are increasingly moving toward unified architectures that integrate visual understanding, image generation, and instruction-guided image editing into a single framework (Deng et al., 2025; Tang et al., 2025). However, these capabilities impose different requirements on visual representations. Visual understanding benefits from high-level semantic abstraction, while image generation requires representations that preserve fine-grained visual details and support efficient modeling and decoding (Ma et al., 2025; Song et al., 2025; Skorokhodov et al., 2025). Visual tokenizers therefore lie at the center of these requirements, as they determine both the visual information exposed to the model and the representation that must be modeled during generation.

Existing approaches mainly follow two design paradigms. The first treats semantic abstraction and content reconstruction of the same image as two separate processing streams, with different encoders dedicated to each function, resulting in substantial information redundancy (Deng et al., 2025). The second paradigm trains a single encoder with both semantic supervision and reconstruction objectives, allowing one representation to serve both purposes (Ma et al., 2025; Song et al., 2025). Although these methods can largely preserve the semantic capabilities of the original VFM, they still fail to match the reconstruction quality achieved by a tokenizer trained solely with a reconstruction objective (Tab. 1). We argue that this does not imply that semantic and reconstruction information cannot coexist in a unified space. Rather, the limited reconstruction capability stems from jointly optimizing the encoder with both semantic and reconstruction objectives, which can impose competing optimization demands on the shared parameters (Song et al., 2025).

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.

Based on this perspective, we introduce **UniFacet**, a unified visual tokenizer that constructs a shared visual space containing both pretrained semantic structure and reconstruction-relevant information. UniFacet keeps a pretrained VFM (Ranzinger et al., 2026; Tschannen et al., 2025) frozen as a semantic anchor and introduces a lightweight residual branch to recover image-specific information that is insufficiently preserved by semantic features. The residual information is added to the semantic features to form a high-dimensional unified representation, which is then projected into a compact and decodable reconstruction facet through a lightweight reconstruction projection. We jointly train the residual branch, reconstruction projection, and decoder using reconstruction supervision. Meanwhile, we apply residual-magnitude regularization to limit the deviation of the unified representation from the pretrained semantic anchor. With this training strategy, UniFacet substantially improves reconstruction fidelity over existing unified visual representation methods while maintaining downstream semantic performance comparable to that of the pretrained VFM, suggesting that missing image-specific information can be recovered without compromising the unified tokenizer's semantic capabilities. The resulting unified representation also serves as source-image conditioning for editing, providing both semantic abstraction and fine-grained visual details.

Although the compact reconstruction facet supports faithful image reconstruction, it is not explicitly optimized for generative learning. Prior work demonstrates that jointly modeling image latents and semantic features can facilitate diffusion training (Kouzelis et al., 2025; 2026; Feng et al., 2026). In their latent-space formulations, image latents are obtained from a reconstruction-oriented VAE, while semantic features are extracted by a separate pretrained VFM. In contrast, UniFacet enriches the pretrained VFM representation with reconstruction-relevant information through residual completion, allowing both reconstruction and semantic facets to be derived from the resulting shared space through lightweight projections. Specifically, We keep the learned reconstruction pathway fixed and jointly train a separate semantic projection with the generative model. This projection extracts a compact semantic facet from the same unified representation, which complements the reconstruction facet to form the generative target. The tokenizer thus provides information-rich source conditioning for image editing and compact, complementary targets for generative modeling.

We instantiate this visual interface in **UniFacet-1.5B**, a unified model for text-to-image generation and instruction-guided image editing. The model uses the shared representation for single-stream source conditioning and jointly models the two compact facets for image synthesis. Experiments show that UniFacet improves reconstruction fidelity while largely preserving pretrained semantic capabilities. With only 1.5B parameters, UniFacet-1.5B achieves 0.89 on GenEval and 7.47 on GEdit-Bench-EN, substantially outperforming the 7B BAGEL model by 0.95 points on GEdit-Bench-EN with approximately one-fifth as many parameters.

Our contributions are summarized as follows:

- We propose **UniFacet**, a unified visual tokenizer that reformulates semantic-reconstruction integration as residual completion of a frozen pretrained semantic representation, enabling high-fidelity reconstruction while largely preserving its semantic structure through regularized residual learning.
- We derive complementary reconstruction and semantic facets from the shared visual space to form a compact generative target, enabling semantic adaptation for generation while preserving the learned reconstruction mapping.
- We develop **UniFacet-1.5B**, a 1.5B-parameter unified model for text-to-image generation and instruction-guided image editing, and validate the proposed visual interface through benchmark evaluations and ablation studies.

2 Related Work

An extended discussion of related work is provided in Appendix D.

Unified multimodal models and visual tokenizers. Unified multimodal models integrate visual understanding, generation, and editing within a unified framework, with visual tokenization as a central design choice. Janus (Wu et al., 2024) and BAGEL (Deng et al., 2025) use separate visual tokenizers for understanding and generation. When both pathways encode the same source image, this design introduces overlapping visual streams. Unified tokenizers instead support understanding

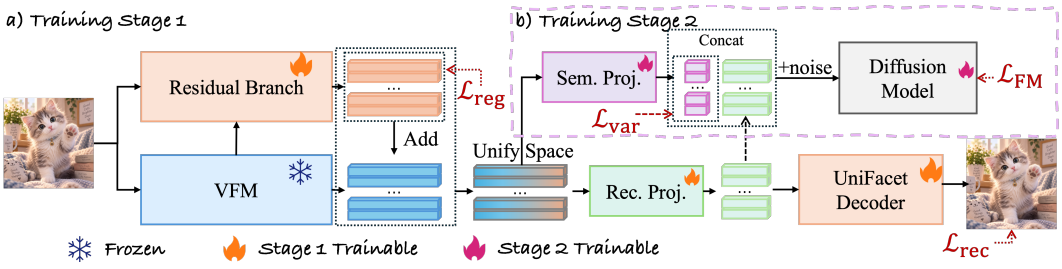

Figure 1: **Training overview of UniFacet.** In Stage 1, the frozen VFM serves as a semantic anchor while the residual branch, reconstruction projection, and image decoder learn reconstruction-relevant information. In Stage 2, the reconstruction pathway is frozen and the semantic projection is adapted jointly with the diffusion model.

and generation through shared visual representations (Xie et al., 2024; Ma et al., 2025; Song et al., 2025; Qu et al., 2024). For continuous visual representations, some studies (Tang et al., 2025; Chen et al., 2026) adapt pretrained semantic encoders for reconstruction through staged training and semantic-preservation constraints. Such adaptation requires balancing semantic preservation and detail recovery, while a reconstruction gap can remain relative to dedicated VAEs.

Semantic representations for generative learning. Semantic features can facilitate diffusion training through alignment or joint modeling. VA-VAE (Yao et al., 2025) introduces semantic alignment into the autoencoder latent space to facilitate generative learning. This approach requires balancing semantic regularization with reconstruction fidelity. A complementary approach jointly models image latents or pixels with separately extracted semantic features, as explored by ReDi (Kouzelis et al., 2025), CoReDi (Kouzelis et al., 2026), and Latent Forcing (Baade et al., 2026). In contrast, UniFacet derives compact reconstruction and semantic facets from a single completed visual representation that already integrates pretrained semantics and reconstruction-relevant details.

3 Method

UniFacet constructs a shared visual space that combines pretrained semantic structure with reconstruction-relevant details through residual completion of a frozen VFM representation. We first describe how to construct this unified visual space and project it into two complementary facets, one focused on reconstruction and the other on semantic structure, providing an effective and compact representation for generative modeling (Sec. 3.1). We then describe how this tokenizer is incorporated into a unified model for image generation and editing, where the shared visual space and its complementary facets are jointly leveraged to improve both source-image conditioning and target-image generation (Sec. 3.2). Fig. 1 summarizes the two-stage training procedure of UniFacet.

3.1 UniFacet: One Visual Space, Complementary Facets

3.1.1 Unified Visual Space and Reconstruction Facet

Residual completion around a frozen semantic anchor. Given an image $x \in \mathbb{R}^{H\times W\times 3}$, a pretrained VFM extracts semantic features

$$z_{\text{vfm}} = E_{\text{vfm}}(x) \in \mathbb{R}^{N\times D}, \tag{1}$$

where N is the number of tokens and D is the feature dimension. We keep E_{vfm} frozen and learn a lightweight residual branch to supplement its features with reconstruction-relevant information:

$$z_{\text{res}} = E_{\text{res}}(x, z_{\text{vfm}}), \quad z_{\text{uni}} = z_{\text{vfm}} + z_{\text{res}} \in \mathbb{R}^{N\times D}. \tag{2}$$

The residual branch receives both image patches and pretrained features, and learns residual image details that are missing from the pretrained semantic representation, denoted as z_{res}. By adding z_{res} to the pretrained semantic features, we obtain z_{uni}, a unified visual representation that retains the semantic structure of the pretrained VFM while complementing it with the fine-grained content details required for reconstruction. Residual addition preserves the spatial token layout and feature dimensionality, producing a single visual sequence.

Compact reconstruction facet. Although z_{uni} retains rich visual information, directly using it as a diffusion target requires modeling a high-dimensional distribution. LV-RAE (Liu et al., 2026) shows that decoders in such spaces can be sensitive to off-manifold deviations in generated latents, leading to visual artifacts. This motivates separating the information-rich unified space from its generative parameterization. We therefore learn a compact reconstruction facet through a lightweight projection that reduces the channel dimension:

$$z_{\text{rec}} = P_{\text{rec}}(z_{\text{uni}}) \in \mathbb{R}^{N \times d_{\text{rec}}}, \quad \hat{x} = \mathcal{D}(z_{\text{rec}}), \quad d_{\text{rec}} \ll D, \tag{3}$$

where z_{rec} is the compact reconstruction facet derived from the unified representation, and $\mathcal{D}$ is the image decoder. The resulting z_{rec} preserves the content information required for faithful reconstruction while using a substantially lower channel dimension d_{rec} than the unified representation dimension D.

Reconstruction learning with constrained residuals. We optimize E_{res}, P_{rec}, and $\mathcal{D}$ using:

$$\mathcal{L}_{\text{tok}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{res}}\mathcal{L}_{\text{reg}}, \quad \mathcal{L}_{\text{rec}} = \|x - \hat{x}\|_1 + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}(x, \hat{x}), \quad \mathcal{L}_{\text{reg}} = \|z_{\text{res}}\|_F^2 \tag{4}$$

The pixel and perceptual losses encourage faithful reconstruction, while the residual penalty limits deviation from the semantic anchor because $\|z_{\text{res}}\|_F^2 = \|z_{\text{uni}} - z_{\text{vfm}}\|_F^2$. Together, the frozen anchor and residual regularization encourage the completed representation to retain pretrained semantic structure while incorporating reconstruction-relevant information. This optimization strategy avoids jointly imposing semantic and reconstruction objectives on the pretrained VFM encoder, and instead reformulates the problem as finding the minimal augmentation needed to equip its existing semantic representation with reconstruction capability. As a result, UniFacet maintains over 99% similarity to the original VFM representation, as measured by CKNNA (Appendix H.3), while achieving substantially higher reconstruction fidelity than existing unified representation methods.

3.1.2 Semantic Facet from the Same Space

Although z_{rec} supports image reconstruction, its training objective does not explicitly encourage a representation suited to generative learning. To preserve the learned reconstruction mapping, we freeze E_{res}, P_{rec}, and $\mathcal{D}$, with E_{vfm} remaining frozen. Then, we introduce a separate projection to extract a compact semantic facet from the unified representation:

$$z_{\text{sem}} = P_{\text{sem},\phi}(z_{\text{uni}}) = \text{BN}\big(\text{Linear}_{\phi}(z_{\text{uni}})\big), \quad z_{\text{sem}} \in \mathbb{R}^{N \times d_{\text{sem}}}. \tag{5}$$

The batch normalization uses no learnable affine parameters and keeps the projected semantic features at a controlled scale to stabilize the subsequent joint training of the semantic projection and diffusion model. We concatenate the spatially aligned facets along the channel dimension,

$$z_{\text{gen}} = \text{Concat}_{\text{ch}}(z_{\text{rec}}, z_{\text{sem}}) \in \mathbb{R}^{N \times (d_{\text{rec}} + d_{\text{sem}})}, \tag{6}$$

to form the generative target. The semantic facet makes semantic information available as an auxiliary prediction target, while the reconstruction facet provides the representation required for decoding. Both facets are derived from the same unified space without an additional visual encoder, and their channel-wise concatenation preserves the original sequence length of N tokens.

Adapting the semantic facet to generation. Following CoReDi (Kouzelis et al., 2026), rather than optimizing z_{sem} toward a fixed semantic anchor, we allow it to evolve jointly with the conditional diffusion model F_θ under the generative objective. This enables the semantic facet to adaptively extract semantic information that is beneficial for generative modeling. Given a target image with generative representation $z = [z_{\text{rec}}; z_{\text{sem}}]$, a conditioning signal c, noise $n = [n_{\text{rec}}; n_{\text{sem}}] \sim \mathcal{N}(0, I)$, and a timestep $t \in [0, 1]$, we construct $z_t = (1 - t)z + tn$ and let the model predict $[\hat{v}_{\text{rec}}; \hat{v}_{\text{sem}}] = F_\theta(z_t, c, t)$. The two facets are supervised with separate velocity targets

$$\mathcal{L}_{\text{FM}}(\theta, \phi) = \mathbb{E}_{x,c,n,t}\Big[\text{MSE}\big(\hat{v}_{\text{rec}},\, n_{\text{rec}} - z_{\text{rec}}\big) + \lambda_{\text{sem}}\ \text{MSE}\big(\hat{v}_{\text{sem}},\, n_{\text{sem}} - \text{sg}(z_{\text{sem}})\big)\Big], \tag{7}$$

where $\text{MSE}(\cdot, \cdot)$ averages squared errors over valid tokens and channels within each facet, $\lambda_{\text{sem}} = 0.25$, and $\text{sg}(\cdot)$ denotes stop-gradient. Gradients to the semantic projection pass through z_t but are blocked through the clean velocity target. This allows the projection to adapt through the prediction

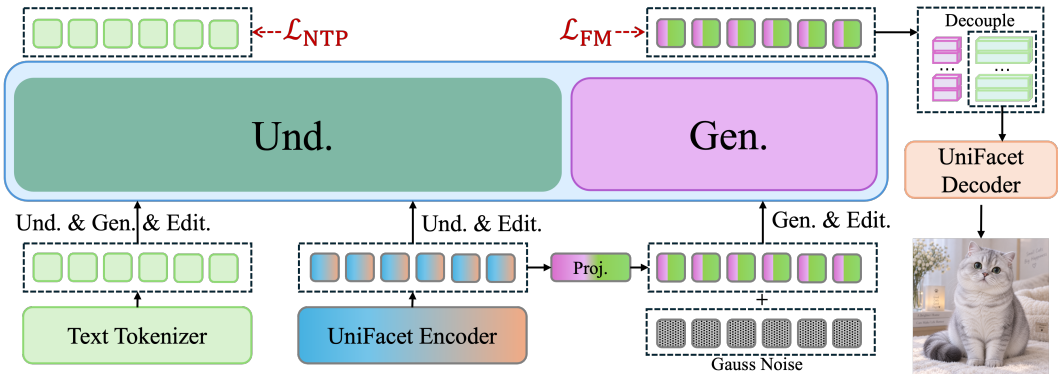

Figure 2: **Overview of UniFacet-1.5B.** The model uses the unified representation as visual context and jointly models the compact reconstruction and semantic facets for target-image generation. At inference, only the generated reconstruction facet is passed to the decoder to synthesize the image.

process without directly changing its target to reduce the loss. We additionally regularize feature variance to discourage collapse:

$$\min_{\theta,\phi} \mathcal{L}_{\text{adapt}} = \mathcal{L}_{\text{FM}}(\theta,\phi) + \lambda_{\text{var}}\mathcal{L}_{\text{var}}(\phi), \tag{8}$$

where $\mathcal{L}_{\text{var}} = \frac{1}{L}\sum_{i=1}^{L} \max\Big(0, \gamma - \sqrt{\text{Var}_{\text{ch}}(z^i_{\text{sem}}) + \epsilon}\Big)$ penalizes the L valid semantic tokens whose standard deviation across channels falls below γ, and ϵ ensures numerical stability. After adaptation, we freeze $P_{\text{sem},\phi}$ and its normalization statistics for all subsequent training. Optimization and freezing schedules are provided in Appendix F.2.

3.2 UniFacet-1.5B

Architecture. We instantiate the visual interface in **UniFacet-1.5B**, a unified model for text-to-image generation and instruction-guided image editing (Fig. 2). Following BAGEL (Deng et al., 2025), we initialize the decoder-only backbone from Qwen2.5-1.5B-Instruct (Qwen Team, 2024) and adopt a MoT architecture with QK-Norm. Understanding and generation tokens are processed by separate Transformer experts and interact through attention over text and visual tokens.

Multimodal training. UniFacet-1.5B is jointly trained for visual understanding, text-to-image generation, and instruction-guided image editing. All tasks read images through z_{uni} and generate images through z_{gen}; they differ only in their conditioning and training targets:

- *Visual understanding:* given the source image representation $z^{\text{src}}_{\text{uni}}$ and a text query q, the model predicts a response $y_{1:T}$.
- *Text-to-image generation:* the conditioning signal c is the text prompt, and the training target is the target image representation $z = z^{\text{tgt}}_{\text{gen}}$.
- *Image editing:* the conditioning signal $c = (u, z^{\text{src}}_{\text{uni}})$ combines the instruction u with the source image representation, and the training target is also $z = z^{\text{tgt}}_{\text{gen}}$.

For visual understanding, we use the autoregressive next-token prediction loss

$$\mathcal{L}_{\text{NTP}} = -\sum_{i=1}^{T} \log p_\theta\big(y_i \mid y_{<i}, q, z^{\text{src}}_{\text{uni}}\big). \tag{9}$$

For text-to-image generation and image editing, we use the same joint flow-matching objective as in Eq. 7, with the visual tokenizer fixed after semantic facet adaptation and only the model parameters θ optimized. Specifically, the target is the fixed generative representation $z^{\text{tgt}}_{\text{gen}} = [z^{\text{tgt}}_{\text{rec}}; z^{\text{tgt}}_{\text{sem}}]$, and the model jointly predicts the velocities of the two facets. Since $P_{\text{sem},\phi}$ is frozen, the stop-gradient in Eq. 7 has no effect, and the variance regularizer is no longer needed. Thus, the overall multimodal training objective is

$$\min_{\theta} \mathcal{L}_{\text{UMM}} = \lambda_{\text{NTP}}\mathcal{L}_{\text{NTP}} + \mathcal{L}_{\text{FM}}(\theta,\phi^*), \tag{10}$$

where ϕ^* denotes the frozen semantic projection and $\lambda_{\text{NTP}} = 0.25$.

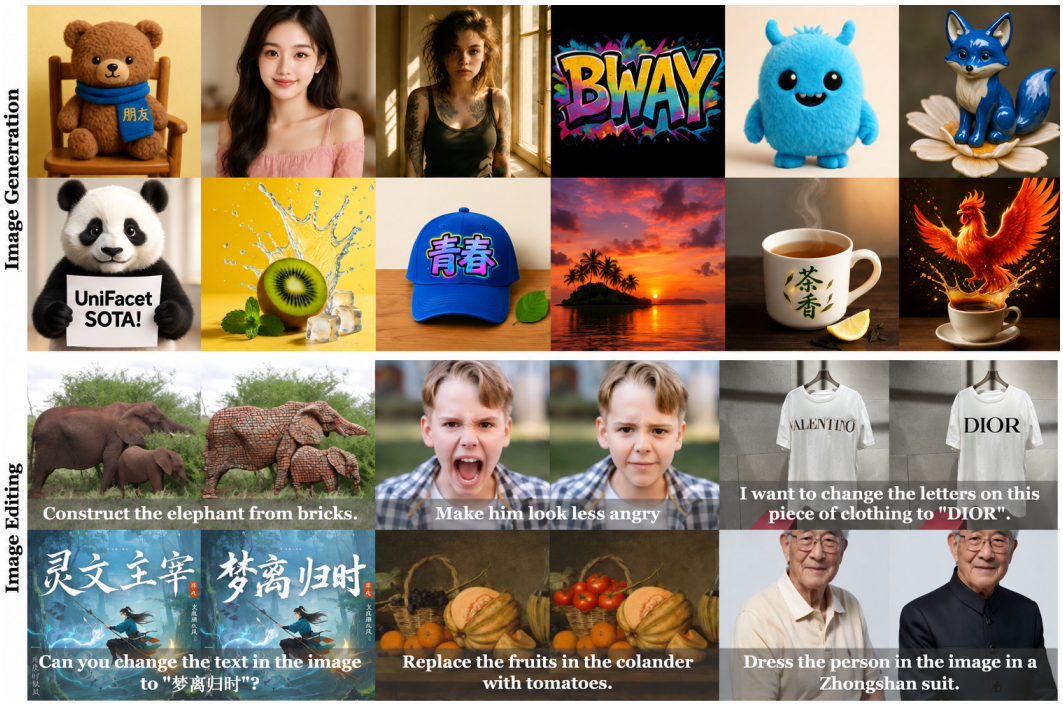

Figure 3: **Qualitative generation and editing results of UniFacet-1.5B.**

Image synthesis at inference. Given a text prompt or an editing condition, UniFacet-1.5B samples the joint representation $\hat{z}_{\text{gen}} = [\hat{z}_{\text{rec}}; \hat{z}_{\text{sem}}]$ starting from Gaussian noise. The semantic facet $\hat{z}_{\text{sem}}$ serves as an auxiliary component during joint generation, while only the reconstruction facet $\hat{z}_{\text{rec}}$ is passed to the image decoder to produce the final image, $\hat{x} = \mathcal{D}(\hat{z}_{\text{rec}})$.

4 Experiments

We evaluate UniFacet on image reconstruction (Section 4.1) and UniFacet-1.5B on text-to-image generation and instruction-based editing (Sections 4.2 and 4.3). We then ablate the key design choices underlying the unified visual space (Section 4.4). Implementation, evaluation, and training details are provided in Appendices E and F.

4.1 Image Reconstruction

We first evaluate whether the unified visual space supports faithful image reconstruction through its reconstruction facet. The full quantitative comparison, qualitative reconstruction comparisons, and CKNNA alignment analysis are provided in Appendix H.

As shown in Table 1, UniFacet with a 64-channel reconstruction facet outperforms all unified tokenizer baselines in both PSNR and SSIM across all three settings, while maintaining strong alignment with the pretrained semantic representation (CKNNA of 0.996; Appendix H.3). On COCO at 512×512, it improves PSNR over UniTok and DualViTok by 6.35 and 8.86 dB, respectively. On ImageNet at 512×512, the corresponding gains are 6.30 and 8.79 dB; this advantage persists at 256×256, with improvements of 5.60 and 8.37 dB. Its reconstruction quality also approaches that of the reconstruction-oriented FLUX-VAE, with a PSNR gap of at most 1.30 dB across all settings. Expanding the reconstruction facet from 64 to 128 channels further improves fidelity, surpassing all listed autoencoders in both PSNR and SSIM across all three settings. However, the 128-channel configuration yields lower GenEval scores under the same training budget, as shown in Section 4.4. We therefore adopt the 64-channel configuration as the default for subsequent experiments to balance reconstruction fidelity and generative performance.

Table 1: **Quantitative comparison of image reconstruction quality.** Within each group, the best result is shown in **bold** and the second best is underlined.

Method	Downsample	Channels	COCO 512		ImageNet 512		ImageNet 256	
			PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Reconstruction-oriented autoencoders								
Qwen-VAE (Wu et al., 2025a)	8×	16	33.60	0.92	32.96	0.92	30.19	0.88
FLUX-VAE (Black Forest Labs, 2024)	8×	16	**34.20**	**0.95**	**33.91**	**0.95**	**30.95**	**0.91**
NextStep-VAE (Team et al., 2025a)	8×	16	33.12	0.93	32.65	0.92	30.07	0.89
Hunyuan-3.0-VAE (Team, 2025)	16×	32	32.32	0.90	31.74	0.89	29.44	0.86
Unified visual tokenizers								
DualToken (Song et al., 2025)	14×	1152	27.06	0.81	26.70	0.79	26.07	0.80
DualViTok (Huang et al., 2025)	16×	64	24.09	0.71	23.82	0.69	21.54	0.61
UniTok (Ma et al., 2025)	16×	64	26.60	0.80	26.31	0.79	24.31	0.73
UniLIP (Tang et al., 2025)	32×	32	25.40	0.74	25.09	0.71	22.97	0.65
UniFacet-RADIOv4 (Ours)	16×	64	32.95	0.93	32.61	0.93	29.91	0.90
UniFacet-RADIOv4 (Ours)	16×	128	**35.95**	**0.97**	**35.77**	**0.97**	**32.60**	**0.95**

Table 2: **Text-to-image generation results on GenEval and DPG-Bench.** Models are grouped into generation-only and unified models, and **bold** and underlined scores indicate the best and second-best results in each column within each group, respectively. † indicates the use of LLM-based prompt rewriting."–" indicates that the corresponding metric was not reported in the original paper.

Models	Params.	GenEval							DPG-Bench					
		1-Obj.	2-Obj.	Count	Colors	Position	Attr.	Overall	Global	Entity	Attribute	Relation	Other	Overall
Generation-only Models														
Emu3-Gen (Wang et al., 2024)	8B	0.98	0.71	0.34	0.81	0.17	0.21	0.54	85.21	86.68	86.84	90.22	83.15	80.60
SDXL (Podell et al., 2023)	3.5B	0.98	0.74	0.39	0.85	0.15	0.23	0.55	83.27	82.43	80.91	86.76	80.41	74.65
DALL-E 3 (Betker et al., 2023)	–	0.96	0.87	0.47	0.83	0.43	0.45	0.67	90.97	89.61	88.39	90.58	89.83	83.50
SD3-Medium (Esser et al., 2024)	2B	**0.99**	**0.94**	0.72	0.89	0.33	0.60	0.74	87.90	91.01	88.83	80.70	88.68	84.08
FLUX.1-dev (Black Forest Labs, 2024)†	12B	0.98	0.93	0.75	**0.93**	0.68	0.65	0.82	74.35	90.00	88.96	90.87	88.33	83.84
Qwen-Image (Wu et al., 2025a)	20B	**0.99**	0.92	**0.89**	0.88	**0.76**	**0.77**	**0.87**	**91.32**	**91.56**	**92.02**	**94.31**	**92.73**	**88.32**
Unified Models														
SEED-X (Ge et al., 2024)	17B	0.97	0.58	0.26	0.80	0.19	0.14	0.49	–	–	–	–	–	–
TokenFlow-XL (Qu et al., 2024)	14B	0.95	0.60	0.41	0.81	0.16	0.24	0.55	–	–	–	–	–	–
Show-o2 (Xie et al., 2025)	7B	**1.00**	0.87	0.58	0.92	0.52	0.62	0.76	89.00	**91.78**	89.96	91.81	**91.64**	**86.14**
Janus-Pro (Chen et al., 2025c)	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80	86.90	88.90	89.40	89.32	89.48	84.19
OmniGen2 (Wu et al., 2025b)	4B	**1.00**	0.95	0.64	0.88	0.55	0.76	0.80	88.81	88.83	90.18	89.37	90.27	83.57
UniWorld-V1 (Lin et al., 2025)	13B	0.99	0.93	0.79	0.89	0.49	0.70	0.80	83.64	88.39	88.44	89.27	87.22	81.38
InternVL-U (Tian et al., 2026)	1.7B	0.99	0.94	0.74	0.91	0.77	0.74	0.85	**90.39**	90.78	90.68	90.29	88.77	85.18
OpenUni-L-1024 (Wu et al., 2025c)	3.1B	0.99	0.92	0.76	0.91	0.82	0.77	0.86	87.01	90.02	89.63	90.28	88.62	83.08
BAGEL (Deng et al., 2025)†	7B	0.98	0.95	**0.84**	**0.95**	0.78	0.77	0.88	88.94	90.37	**91.29**	90.82	88.67	85.07
Mogao (Liao et al., 2025)	7B	**1.00**	**0.97**	0.83	0.93	**0.84**	0.80	**0.89**	82.37	90.03	88.26	93.18	85.40	84.33
UniFacet-1.5B (Ours)	**1.5B**	**1.00**	0.96	**0.84**	0.93	0.80	**0.83**	**0.89**	81.99	90.91	89.06	**93.95**	82.70	85.15

4.2 Text-to-Image Generation

We evaluate text-to-image generation on GenEval (Ghosh et al., 2023) and DPG-Bench (Hu et al., 2024), with results in Table 2. On GenEval, UniFacet-1.5B achieves 0.89, tying the highest overall score in the comparison with the 7B Mogao model. Its clearest advantage is attribute binding, where it reaches 0.83, compared with 0.80 for Mogao and 0.77 for BAGEL. Together with a two-object score of 0.96, this indicates strong performance in composing multiple objects and assigning their attributes. On DPG-Bench, UniFacet achieves an overall score of 85.15 with only 1.5B parameters, slightly surpassing the 7B BAGEL model (85.07) with approximately one-fifth as many parameters. UniFacet is particularly strong in relation modeling, achieving the highest Relation score among the listed unified models (93.95) and outperforming BAGEL by 3.13 points. We provide qualitative generation results in Figure 3, with additional examples in Appendix J.

4.3 Image Editing

We evaluate instruction-based image editing on ImgEdit (Ye et al., 2025) and the English and Chinese sets of GEdit-Bench (Liu et al., 2025), with results summarized in Table 3. On ImgEdit, UniFacet-1.5B achieves an overall score of 4.03, outperforming Ovis-U1 (3.97), InternVL-U (3.67), and BAGEL (3.20). Compared with the 7B BAGEL model, UniFacet improves the overall score by 0.83 points with approximately one-fifth as many parameters. The full category breakdown in Appendix K reveals particularly large gains in Extract (+1.94) and Remove (+1.56), highlighting its effectiveness in these editing operations. On GEdit-Bench, UniFacet achieves the highest semantic consistency (SC) scores among the listed models in both English and Chinese. On the English set, it obtains SC of 8.44, perceptual quality (PQ) of 7.06, and an overall score of 7.47, improving over BAGEL by 1.08 points in SC and 0.23 points in PQ. Its SC also exceeds that of Z-Image-Edit and

Table 3: **Image editing results on ImgEdit and GEdit-Bench (EN/CN). Bold** and underlined scores denote the best and second-best results across all listed models, respectively."–" indicates that the corresponding metric was not reported in the original paper.

Model	# Params	ImgEdit					GEdit-Bench-EN			GEdit-Bench-CN		
		Add	Remove	Replace	Style	Overall↑	SC↑	PQ↑	O↑	SC↑	PQ↑	O↑
Closed-source Models												
FLUX.1-Kontext [Pro] (Labs et al., 2025)	–	4.25	3.57	4.56	4.57	4.00	7.02	7.60	6.56	–	–	–
GPT-Image-1 [High] (OpenAI, 2025b)	–	4.61	3.66	4.35	**4.93**	4.20	7.85	7.62	7.53	7.67	7.56	7.30
Nano-Banana (Google, 2025)	–	**4.62**	**4.39**	4.34	4.18	4.29	7.40	**8.45**	7.29	7.51	**8.31**	7.25
Open-source Models												
MagicBrush (Zhang et al., 2023)	0.9B	2.84	1.58	1.97	2.38	1.83	4.68	5.66	4.52	–	–	–
Instruct-Pix2Pix (Brooks et al., 2022)	0.9B	2.45	1.50	2.01	3.55	1.88	3.58	5.49	3.68	–	–	–
AnyEdit (Yu et al., 2024)	1.3B	3.18	2.23	2.47	2.85	2.45	3.18	5.82	3.21	–	–	–
UltraEdit (Zhao et al., 2024)	2B	3.44	1.45	2.96	3.76	2.70	–	–	4.83	–	–	–
OmniGen (Xiao et al., 2024)	3.8B	3.47	2.43	2.94	4.19	2.96	5.96	5.89	5.06	–	–	–
ICEdit (Zhang et al., 2025b)	12B	3.58	2.93	3.15	3.84	3.05	5.11	6.85	4.84	–	–	–
Step1X-Edit (Liu et al., 2025)	12B	3.88	2.41	3.40	4.63	3.06	7.66	7.35	6.97	7.20	6.87	6.86
BAGEL (Deng et al., 2025)	7B	3.56	2.62	3.30	4.49	3.20	7.36	6.83	6.52	7.34	6.85	6.50
UniWorld-V1 (Lin et al., 2025)	12B	3.82	3.24	3.47	4.21	3.26	4.93	7.43	4.85	–	–	–
OmniGen2 (Wu et al., 2025b)	4B	3.57	3.20	3.74	4.81	3.44	7.16	6.77	6.41	–	–	–
InternVL-U (Tian et al., 2026)	1.7B	4.13	3.39	4.13	4.77	3.67	–	–	6.66	–	–	–
FLUX.1-Kontext [Dev] (Labs et al., 2025)	12B	4.12	3.09	4.22	4.51	3.71	6.52	7.38	6.00	–	–	–
Ovis-U1 (Wang et al., 2025)	1.2B	3.99	4.15	4.38	4.86	3.97	–	–	6.42	–	–	–
Qwen-Image-Edit (Wu et al., 2025a)	20B	4.38	4.14	**4.66**	4.81	4.27	8.00	7.86	7.56	7.82	7.79	7.52
Z-Image-Edit (Team et al., 2025b)	6B	4.40	4.13	4.57	4.85	**4.30**	8.11	7.72	**7.57**	8.03	7.80	7.54
UniFacet-1.5B (Ours)	1.5B	4.21	4.18	4.31	4.49	4.03	**8.44**	7.06	7.47	**8.61**	7.00	**7.56**

Table 4: **Semantic preservation after residual completion.** The listed representation is used consistently during training and inference. Avg. is the arithmetic mean over the six benchmarks.

Anchor	Feature	SEED	MMStar	MMB	OCR	RWQA	AI2D	Avg.↑	Δ Avg.
RADIOv4	z_{vfm}	76.1	55.0	73.6	63.0	61.2	**78.1**	67.83	–
RADIOv4	z_{uni}	**76.2**	**55.3**	**74.7**	**64.2**	**61.7**	77.6	**68.28**	**+0.45**
SigLIP2	z_{vfm}	**76.0**	54.8	**74.0**	**53.8**	62.0	**78.9**	66.58	–
SigLIP2	z_{uni}	75.9	**55.5**	73.1	53.6	**63.4**	78.6	**66.68**	**+0.10**

Qwen-Image-Edit. This strength extends to the Chinese set, where UniFacet leads in both SC (8.61) and overall score (7.56), with PQ of 7.00. These results highlight strong semantic consistency across both language settings at the 1.5B model scale. We further examine the contribution of the unified source representation in Section 4.4. We provide qualitative editing results in Figure 3 (bottom), with additional examples in Appendix M.

4.4 Ablation Studies

We conduct three ablation studies to examine the key design choices of UniFacet. Unless otherwise specified, each comparison changes only the component under study while keeping the remaining architecture and training recipe fixed. Detailed training protocols are provided in Appendix F.4.

Does residual completion preserve pretrained semantics? A central goal of UniFacet is to enrich a pretrained semantic representation with reconstruction-relevant information without disrupting its original semantic organization. We evaluate this property using RADIOv4-H (Ranzinger et al., 2026) and SigLIP2 (Tschannen et al., 2025) as semantic anchors. All understanding models in this comparison use Qwen2.5-1.5B-Instruct (Qwen Team, 2024). For each anchor, we train and evaluate the understanding model consistently using either the original VFM representation z_{vfm} or the completed unified representation z_{uni}. As shown in Table 4, replacing z_{vfm} with z_{uni} does not degrade visual understanding and instead yields a small overall improvement. With RADIOv4-H, the average score increases by 0.45 points, with improvements on five of the six benchmarks, including gains of 1.2 points on OCRBench and 1.1 points on MMBench. With SigLIP2, the average score remains nearly unchanged, with a slight gain of 0.10 points. We further perform cross-representation evaluation by swapping z_{vfm} and z_{uni} only at inference time. The average score changes by at most 0.18 points, with full results provided in Appendix G. CKNNA scores of 0.995–0.996 on 500 COCO 2017 validation images at 512×512 resolution further indicate strong alignment between z_{vfm} and z_{uni} (Appendix H.3). These results indicate that residual completion introduces additional reconstruction-relevant information while largely preserving the semantic coordinate system of the pretrained representation.

Do complementary facets facilitate generative learning? UniFacet derives reconstruction and semantic facets from the same unified visual space and jointly models them as the generative target.

To isolate the contribution of the semantic facet, we first fix the reconstruction-facet dimension and vary the dimensionality of z_{sem}. As shown in Table 5, augmenting z_{rec} with the semantic facet consistently improves GenEval from 0.65 to 0.70–0.71. The improvement is relatively insensitive to the exact semantic dimensionality, suggesting that the main benefit comes from exposing an explicit semantic coordinate rather than simply increasing latent capacity. We observe an even larger effect when increasing the capacity of the reconstruction facet. Using the larger z_{rec} alone degrades GenEval to 0.57, whereas introducing the semantic facet improves the score to 0.66, corresponding to a gain of 0.09. Because the image decoder consumes only z_{rec}, the semantic facet does not increase the reconstruction capacity of the decoder. Instead, it provides complementary semantic structure that makes the joint generative representation easier to learn. These results support the multi-facet design of UniFacet: the unified space preserves rich visual information, while compact complementary facets expose a more generation-friendly representation.

Table 5: **Effect of complementary generative facets.** We vary the semantic-facet dimension under different reconstruction-facet capacities. Adding the semantic facet consistently improves GenEval.

Generative Target	d_{rec}	d_{sem}	GenEval↑	Δ
z_{rec}	64	–	0.65	–
$(z_{\text{rec}}, z_{\text{sem}})$	64	8	0.70	+0.05
		16	0.70	+0.05
		32	**0.71**	**+0.06**
z_{rec}	128	–	0.57	–
$(z_{\text{rec}}, z_{\text{sem}})$	128	16	**0.66**	**+0.09**

Does the UniFacet tokenizer improve image editing? We compare UniFacet with two reproduced BAGEL variants that share the same backbone, training data, schedule, and evaluation protocol, and differ only in the tokenizer setup.

BAGEL uses ViT-only or ViT+VAE source conditioning, while UniFacet conditions on the completed representation z_{uni}. With N tokens per source stream, UniFacet requires N source tokens compared with $2N$ for dual-stream BAGEL. On GEdit-Bench-EN (Table 6), adding the VAE stream to BAGEL raises the overall score from 5.92 to 6.29, with a larger improvement in perceptual quality (PQ, $+0.60$) than semantic consistency (SC, $+0.11$). UniFacet further improves the overall score to 7.35 and achieves the highest SC and PQ among these configurations while using a single source stream. These results support the effectiveness of UniFacet's tokenizer configuration as a whole. We then examine whether the trained UniFacet model relies on its completed source representation by replacing z_{uni} with z_{vfm} only at inference, leaving the model unchanged. The overall score drops from 7.35 to 3.78, showing that the original VFM features cannot directly replace the completed representation for editing. In contrast to the small changes observed in the understanding swap experiments (Appendix G), this result indicates that compatibility in visual understanding does not imply interchangeability for editing. Because the swap also introduces a mismatch between training and inference representations, it does not isolate the contribution of residual information alone.

Table 6: **Editing ablation on GEdit-Bench-EN.** † replaces z_{uni} with z_{vfm} only at inference, without retraining.

Configuration	Source Tokens	SC↑	PQ↑	Overall↑
BAGEL (ViT only)	N	7.64	4.95	5.92
BAGEL (ViT + VAE)	$2N$	7.75	5.55	6.29
UniFacet (z_{uni})	N	**8.35**	**6.88**	**7.35**
UniFacet (z_{vfm})†	N	5.46	2.85	3.78

5 Conclusion

We presented UniFacet, a unified visual tokenizer that learns reconstruction-relevant information through regularized residual completion around a frozen pretrained semantic anchor. A separate semantic projection then adapts to generative learning while the reconstruction pathway remains fixed. The resulting shared space provides single-stream source conditioning for image editing, while its complementary compact facets serve as joint generative targets. Experiments demonstrate strong reconstruction fidelity with largely preserved semantic capabilities, and UniFacet-1.5B achieves competitive text-to-image generation and instruction-guided editing, including 0.89 on GenEval and 7.56 on GEdit-CN. These results support a unified visual interface that preserves rich image information while providing distinct, compact representations for generative learning.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Alan Baade, Eric Ryan Chan, Kyle Sargent, Changan Chen, Justin Johnson, Ehsan Adeli, and Fei-Fei Li. Latent forcing: Reordering the diffusion trajectory for pixel-space image generation. *arXiv preprint arXiv:2602.11401*, 2026.

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

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

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. COYO-700M: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

Bowei Chen, Sai Bi, Hao Tan, He Zhang, Tianyuan Zhang, Zhengqi Li, Yuanjun Xiong, Jianming Zhang, and Kai Zhang. AlignTok: Aligning visual foundation encoders to tokenizers for diffusion models. In *International Conference on Learning Representations*, 2026. URL `https://arxiv.org/abs/2509.25162`.

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. *arXiv preprint arXiv:2505.09568*, 2025a.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. ShareGPT-4o-Image: Aligning multimodal models with GPT-4o-level image generation. *arXiv preprint arXiv:2506.18095*, 2025b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025c.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

Ruoyu Feng, Jinming Liu, Yuqi Wang, Xin Cheng, Boyuan Liu, Shanglin Li, Hanshen Zhu, Wenfeng Lin, Mingyu Guo, and Xin Jin. SeFi-Image: A text-to-image foundation model with semantic-first diffusion. *arXiv preprint arXiv:2606.22568*, 2026.

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

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yue Gong, Hongyu Li, Shanyuan Liu, Bo Cheng, Yuhang Ma, Liebucha Wu, Xiaoyu Wu, Manyuan Zhang, Dawei Leng, Yuhui Yin, and Lijun Zhang. RPiAE: A representation-pivoted autoencoder enhancing both image generation and editing. *arXiv preprint arXiv:2603.19206*, 2026.

Google. Introducing Gemini 2.5 Flash Image. `https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/`, 2025.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253. Association for Computational Linguistics, 2020.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Runhui Huang, Chunwei Wang, Junwei Yang, Guansong Lu, Yunlong Yuan, Jianhua Han, Lu Hou, Wei Zhang, Lanqing Hong, Hengshuang Zhao, et al. ILLUME+: Illuminating unified MLLM with dual visual tokenization and diffusion refinement. *arXiv preprint arXiv:2504.01934*, 2025.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 448–456. PMLR, 2015.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, 2016.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Theodoros Kouzelis, Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Boosting generative image modeling via joint image-feature synthesis. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Theodoros Kouzelis, Spyros Gidaris, and Nikos Komodakis. Coevolving representations in joint image-feature diffusion. *arXiv preprint arXiv:2604.17492*, 2026.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 2020.

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

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking multimodal LLMs with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

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

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. UniWorld-V1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1X-Edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Siyu Liu, Chujie Qin, Hubery Yin, Qixin Yan, Zheng-Peng Duan, Chen Li, Jing Lyu, Chun-Le Guo, and Chongyi Li. Improving reconstruction of representation autoencoder. *arXiv preprint arXiv:2602.08620*, 2026.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023a.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. UniTok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025.

OpenAI. Introducing GPT-4.1 in the API, 2025a. URL https://openai.com/index/gpt-4-1/.

OpenAI. GPT Image 1: State-of-the-art image generation model. https://platform.openai.com/docs/models/gpt-image-1, 2025b.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K. Du, Zehuan Yuan, and Xinglong Wu. TokenFlow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024.

Qwen Team. Qwen2.5: A party of foundation models, 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Mike Ranzinger, Greg Heinrich, Collin McCarthy, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. C-RADIOv4 (tech report). *arXiv preprint arXiv:2601.17237*, 2026.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.15301.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. *arXiv preprint arXiv:2502.14831*, 2025.

Wei Song, Yuran Wang, Zijia Song, Yadong Li, Zenan Zhou, Long Chen, Jianhua Xu, Jiaqi Wang, and Kaicheng Yu. DualToken: Towards unifying visual understanding and generation with dual visual vocabularies. *arXiv preprint arXiv:2503.14324*, 2025.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Hao Tang, Chenwei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. UniLiP: Adapting CLIP for unified multimodal understanding, generation and editing. *arXiv preprint arXiv:2507.23278*, 2025.

NextStep Team, Chunrui Han, Guopeng Li, Jingwei Wu, Quan Sun, Yan Cai, Yuang Peng, Zheng Ge, Deyu Zhou, Haomiao Tang, et al. NextStep-1: Toward autoregressive image generation with continuous tokens at scale. *arXiv preprint arXiv:2508.10711*, 2025a.

Tencent Hunyuan Foundation Model Team. HunyuanImage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, et al. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025b.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, et al. InternVL-U: Democratizing unified multimodal models for understanding, reasoning, generation and editing. *arXiv preprint arXiv:2603.09877*, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, et al. Ovis-U1 technical report. *arXiv preprint arXiv:2506.23044*, 2025.

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

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. FineVision: Open data is all you need. *arXiv preprint arXiv:2510.17269*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

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.

Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. OpenUni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*, 2025c.

xAI. RealWorldQA, 2024. URL https://huggingface.co/datasets/xai-org/RealworldQA.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024.

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. *arXiv preprint arXiv:2408.12528*, 2024.

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

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712. IEEE, 2025.

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

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. AnyEdit: Mastering unified high-quality image editing for any idea. *arXiv preprint arXiv:2411.15738*, 2024.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. MagicBrush: A manually annotated dataset for instruction-guided image editing. *arXiv preprint arXiv:2306.10012*, 2023.

Shilong Zhang, He Zhang, Zhifei Zhang, Chongjian Ge, Shuchen Xue, Shaoteng Liu, Mengwei Ren, Soo Ye Kim, Yuqian Zhou, Qing Liu, Daniil Pakhomov, Kai Zhang, Zhe Lin, and Ping Luo. Both semantics and reconstruction matter: Making representation encoders ready for text-to-image generation and editing. *arXiv preprint arXiv:2512.17909*, 2025a.

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

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. UltraEdit: Instruction-based fine-grained image editing at scale. *arXiv preprint arXiv:2407.05282*, 2024.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026.

APPENDIX CONTENTS

A AI Assistance **16**

B Impact Statement **16**

C Limitations **16**

D Extended Related Work **16**

D.1 Unified Multimodal Models and Visual Tokenizers 16

D.2 Representation Autoencoders 17

D.3 Semantic Representations for Generative Learning 17

E Implementation Details **17**

E.1 Architecture 17

E.2 Evaluation Protocol 18

F Training Details **19**

F.1 Tokenizer Training 19

F.2 Semantic-Facet Adaptation 19

F.3 UniFacet-1.5B Training 20

F.4 Ablation Training Details 20

G Cross-Representation Compatibility **22**

H Additional Reconstruction and Alignment Results **22**

H.1 Full Quantitative Comparison 22

H.2 Qualitative Reconstruction Comparisons 23

H.3 CKNNA Alignment 24

I Direct Semantic Alignment of the Reconstruction Latent **25**

J Additional Text-to-Image Generation Results **26**

K Full ImgEdit Results **28**

L Discussion of Perceptual Quality on GEdit-Bench **29**

M Additional Image Editing Results **29**

A AI ASSISTANCE

We used large language models (LLMs) as assistive tools to polish the manuscript, search for and identify related papers, and generate some non-critical code for our experiments. We did not use LLMs to generate synthetic datasets or prove mathematical claims. All LLM-assisted work, including the retrieved references and generated code, was reviewed and checked by the authors. The authors take full responsibility for the final content of this paper, including its text, references, methodology, results, and conclusions.

B IMPACT STATEMENT

This paper advances unified visual representations for image reconstruction, text-to-image generation, and instruction-guided image editing, with an emphasis on reconstruction fidelity and preservation of pretrained semantic structure. These advances support applications in creative content generation and generative AI research. Our method improves representation learning within existing generative workflows and shares the ethical considerations and societal challenges associated with large-scale generative systems, including potential misuse, biases in generated content, and privacy concerns.

C LIMITATIONS

Reconstruction fidelity. Although UniFacet substantially improves reconstruction quality over existing unified visual tokenizers, its default 64-channel configuration still exhibits a small fidelity gap relative to leading reconstruction-oriented VAEs such as FLUX-VAE (Table 1). Further narrowing this gap while retaining a compact representation for generative modeling remains an important direction for future work.

Model scale and output resolution. Due to limited computational resources, we have not explored larger generative backbones or higher-resolution image generation and editing. Our full system uses a Qwen-based 1.5B backbone, with the longest image side capped at 512 pixels during generation and editing training. Evaluating UniFacet at larger model scales and higher output resolutions is left for future work.

D EXTENDED RELATED WORK

D.1 UNIFIED MULTIMODAL MODELS AND VISUAL TOKENIZERS

Unified multimodal architectures combine visual understanding and generation while differing in how visual information is encoded and modeled. Janus (Wu et al., 2024) separates understanding and generation encoders while retaining a shared autoregressive Transformer. Show-o (Xie et al., 2024) combines autoregressive text modeling with discrete image diffusion and uses a shared discrete visual tokenizer in its default formulation. BAGEL (Deng et al., 2025) connects understanding and generation experts through shared multimodal attention while using separate ViT and VAE encoders. UniFacet-1.5B follows BAGEL's Mixture-of-Transformer-Experts architecture, but derives source-image context and generative targets from a shared visual space.

At the tokenizer level, discrete approaches coordinate semantic and reconstruction information through supervision and codebook design. UniTok (Ma et al., 2025) combines reconstruction and contrastive supervision with multi-codebook quantization to increase representation capacity. DualToken (Song et al., 2025) applies reconstruction and semantic supervision at different encoder depths and constructs separate visual vocabularies, while TokenFlow (Qu et al., 2024) couples semantic and pixel codebooks through shared token indices. For continuous representations, UniLiP (Tang et al., 2025) fine-tunes a pretrained semantic encoder for reconstruction using staged training and self-distillation. AlignTok (Chen et al., 2026) first trains an adapter and decoder with the encoder frozen, then jointly adapts the components with a semantic-preservation loss, and finally refines the decoder. These approaches seek to retain pretrained semantic structure while recovering the fine-grained information required for reconstruction.

D.2 Representation Autoencoders

Representation autoencoders were introduced to facilitate generative modeling by replacing conventional VAE latents with semantically structured features from pretrained visual encoders (Zheng et al., 2026). However, these features can lack image-specific details, limiting reconstruction fidelity (Liu et al., 2026; Gong et al., 2026). This limitation has motivated efforts to improve reconstructability while retaining pretrained semantics. RPiAE (Gong et al., 2026) fine-tunes a representation encoder under semantic-preservation regularization, while SVG (Shi et al., 2026) and LV-RAE (Liu et al., 2026) supplement frozen semantic features with missing low-level information. In particular, LV-RAE learns a lightweight residual encoder and regularizes the augmented representation toward the semantic anchor, using the resulting high-dimensional features as both diffusion latents and decoder inputs. It further addresses decoder sensitivity to latent perturbations through noise-augmented decoder fine-tuning and controlled noise injection at inference.

UniFacet shares residual completion with LV-RAE, but addresses a different design problem. LV-RAE focuses on improving reconstruction fidelity and decoding robustness in pretrained representation space, whereas UniFacet addresses the different representational demands of image conditioning and generative modeling. Its central distinction is to unify the visual information source without requiring conditioning and generation to use the same representation, allowing rich image context and compact generative modeling to coexist.

D.3 Semantic Representations for Generative Learning

Semantic information can aid generative learning through latent alignment or joint prediction. VA-VAE (Yao et al., 2025) introduces semantic alignment into the autoencoder latent space, requiring semantic regularization to be balanced with reconstruction fidelity. Joint modeling instead makes semantic features an explicit generative target alongside image content. ReDi (Kouzelis et al., 2025) jointly models VAE latents and compressed semantic features, while CoReDi (Kouzelis et al., 2026) further adapts the semantic projection together with the diffusion model. Latent Forcing (Baade et al., 2026) jointly processes semantic latents and pixels with separate noise schedules, allowing semantic denoising to precede detail synthesis. These approaches obtain semantic targets from an additional pretrained visual encoder. Following CoReDi's joint adaptation strategy, UniFacet learns a compact semantic projection with the generative model while keeping its reconstruction pathway fixed. Both generative facets come from the same completed visual space, so semantic targets require no additional visual encoder or direct alignment constraint on the reconstruction facet.

E Implementation Details

E.1 Architecture

E.1.1 Tokenizer Architecture

UniFacet keeps the pretrained vision foundation model frozen as a semantic anchor and learns a lightweight residual branch that supplies the complementary visual information required for reconstruction. Unless otherwise noted, we use RADIOv4-H (Ranzinger et al., 2026), specifically the C-RADIOv4-H checkpoint,[1] as the semantic anchor. For the ablations explicitly using SigLIP2 (Tschannen et al., 2025), we use the `siglip2-so400m-patch16-512` checkpoint.[2] The spatial feature dimension is $D = 1280$ for RADIOv4-H and $D = 1152$ for SigLIP2. The tokenizer operates at a spatial downsampling factor of $16\times$, and we report two reconstruction-facet capacities, $d_{\text{rec}} \in \{64, 128\}$, with $d_{\text{rec}} = 64$ as the default.

Residual encoder. The residual encoder consists of six Transformer blocks (Vaswani et al., 2017) with a hidden dimension of 1,280, 20 attention heads, and an FFN expansion ratio of 4. It takes embeddings of non-overlapping 16×16 image patches together with the frozen VFM features, with input projections aligning their channel dimensions to the encoder width. The encoder uses rotary position embeddings (RoPE) (Su et al., 2021) and gated feed-forward layers (Shazeer, 2020) with

[1] `https://huggingface.co/nvidia/C-RADIOv4-H`

[2] `https://huggingface.co/google/siglip2-so400m-patch16-512`

GELU activations (Hendrycks & Gimpel, 2016). A final LayerNorm (Ba et al., 2016) and linear projection map the encoder outputs to the VFM feature dimension D. The output projection is initialized to zero, so the unified representation initially matches the pretrained semantic anchor.

Reconstruction projector. The reconstruction projector P_{rec} maps z_{uni} to the compact reconstruction facet z_{rec} while preserving the spatial token layout. Its input is projected to 1,280 channels when needed, followed by four Transformer blocks with 20 attention heads and an FFN expansion ratio of 4. A linear layer then reduces the channel dimension to d_{rec}, followed by LayerNorm.

Image decoder. The decoder first projects each reconstruction token from d_{rec} to 768 channels, then applies 12 Transformer blocks with 12 attention heads and an FFN expansion ratio of 4. These blocks use rotary position embeddings, followed by a final LayerNorm and an output head that converts the tokens into 16×16 RGB patches to reconstruct the image. The reconstruction projector and decoder use the same gated feed-forward design as the residual encoder, and all three components use LayerNorm with $\epsilon = 10^{-5}$. Only z_{rec} is required for image decoding.

Semantic projector. The learnable semantic projector is a single linear layer that maps each unified visual token from D to d_{sem} channels, with $d_{\mathrm{sem}} = 16$ by default. Its output is normalized by batch normalization (Ioffe & Szegedy, 2015) without learnable affine parameters to obtain z_{sem}, consistent with the definition in the main text. Concatenating z_{rec} and z_{sem} along the channel dimension yields the generative target z_{gen} without changing the spatial token layout. The main UniFacet-1.5B model uses RADIOv4-H with $(D, d_{\mathrm{rec}}, d_{\mathrm{sem}}) = (1280, 64, 16)$, giving $d_{\mathrm{rec}} + d_{\mathrm{sem}} = 80$ channels per spatial token and $z_{\mathrm{gen}} \in \mathbb{R}^{N \times 80}$. Although increasing d_{sem} from 16 to 32 at $d_{\mathrm{rec}} = 64$ slightly improves GenEval from 0.70 to 0.71 (Table 5), we retain $d_{\mathrm{sem}} = 16$ as the default for consistency across experiments. The other facet-dimension combinations in Table 5 are used for ablation experiments.

E.1.2 UniFacet-1.5B Architecture

UniFacet-1.5B follows the BAGEL framework (Deng et al., 2025) for unified multimodal modeling. We initialize its decoder-only language backbone from Qwen2.5-1.5B-Instruct (Qwen Team, 2024) and augment it with query–key (QK) normalization (Henry et al., 2020) and a Mixture-of-Transformer-Experts (MoT) architecture to support autoregressive language modeling and continuous visual generation. For visual understanding, the model consumes z_{uni} as visual context for language prediction. For text-to-image generation, it predicts velocities over the concatenated reconstruction and semantic facets, $z_{\mathrm{gen}} = [z_{\mathrm{rec}}; z_{\mathrm{sem}}]$, conditioned on the text prompt. For image editing, the source image is represented by z_{uni}, and the model predicts the target image in the compact z_{gen} space conditioned on the source representation and editing instruction. At inference time, both facets are jointly sampled, and the reconstruction facet is passed to the frozen image decoder to obtain the output image.

E.2 Evaluation Protocol

Image reconstruction. We evaluate reconstruction on the COCO 2017 validation set (Lin et al., 2014) and on 5,000 images sampled from the ImageNet validation set (Russakovsky et al., 2015). We report peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) (Wang et al., 2004) at 512×512 on both datasets, with an additional evaluation at 256×256 on ImageNet. For tokenizers with a fixed input resolution, we adapt the evaluation procedure to their native resolution, for example by reconstructing a 512×512 image using four 256×256 patches.

Generation and editing. Text-to-image generation is evaluated on DPG-Bench (Hu et al., 2024) and GenEval (Ghosh et al., 2023). Instruction-based editing is evaluated on ImgEdit (Ye et al., 2025) and the English and Chinese sets of GEdit-Bench (Liu et al., 2025). We evaluate both editing benchmarks on their full sets using GPT-4.1 (OpenAI, 2025a) as the judge.

Visual understanding. Visual understanding in the ablations is evaluated on SEED-Bench (Li et al., 2023), MMStar (Chen et al., 2024), MMBench (Liu et al., 2023a), OCRBench (Liu et al., 2023b), RealWorldQA (xAI, 2024), and AI2D (Kembhavi et al., 2016).

F TRAINING DETAILS

This section describes tokenizer training, semantic-facet adaptation, and downstream UniFacet-1.5B training. Stage 1 in Fig. 1 corresponds to tokenizer training, which comprises Phases T1 and T2 below; Stage 2 corresponds to semantic-facet adaptation. The downstream UniFacet-1.5B curriculum uses the distinct identifiers C1–C4, followed by supervised fine-tuning (SFT), with the learned visual interface held fixed.

F.1 TOKENIZER TRAINING

The tokenizer is trained in two phases, T1 and T2, with the vision foundation model frozen throughout so that gradients reach only the residual branch, the reconstruction projector, and the decoder. Phase T1 establishes reconstruction at a fixed resolution of 256×256 on ImageNet. Phase T2 continues on a larger internal corpus at mixed resolutions, where we preserve the native aspect ratio of each image and cap the total number of pixels at 1024×1024, so that a single tokenizer serves the varying resolutions required by generation and editing. Both phases minimize $\mathcal{L}_{\text{tok}}$ in Eq. 4 with $\lambda_{\text{LPIPS}} = 1.0$ and $\lambda_{\text{res}} = 2.5$, optimized with AdamW using zero weight decay, a constant learning-rate schedule with 500 warmup steps, and bfloat16 mixed precision. Table 7 lists the remaining settings for T1 and T2.

Table 7: **Tokenizer training configurations for Phases T1 and T2.**

	Phase T1	Phase T2
Training data	ImageNet	Internal corpus
Resolution	256×256	Native aspect ratio, $\leq 1024 \times 1024$
Training steps	100k	50k
Global batch size	1536	96
Learning rate	2×10^{-4}	2×10^{-5}
Optimizer	AdamW (Loshchilov & Hutter, 2017)	AdamW (Loshchilov & Hutter, 2017)
Grad. Clip	1.0	1.0
Betas	(0.9, 0.999)	(0.9, 0.999)
Schedule	Constant	Constant
Warmup steps	500	500
Precision	bfloat16	bfloat16

F.2 SEMANTIC-FACET ADAPTATION

The semantic projector used by UniFacet-1.5B is learned during the joint-training run for the complementary generative-facet ablation (Table 5; Appendix F.4). The 30k-step semantic adaptation described here is part of that 50k-step ablation run and is completed before curriculum phase C1 begins.

Trainable components. We adapt the semantic facet to the generative objective while preserving the pretrained reconstruction pathway. The visual backbone, residual branch, reconstruction projector, and image decoder remain frozen, so the reconstruction facet z_{rec} is fixed throughout semantic-facet adaptation. The semantic projector is jointly optimized with the visual generative model, allowing z_{sem} to adapt to continuous visual prediction.

Training objective. We concatenate the reconstruction and semantic facets to form the generative target z_{gen} and optimize the velocity-prediction objective with variance regularization in Eq. 8. The clean representation used to construct the velocity target is detached, while the noisy input remains differentiable with respect to the semantic projector. Affine-free batch normalization stabilizes the scale of the projected features, and the variance regularizer discourages collapse of the semantic facet.

Optimization and freezing. We use AdamW (Loshchilov & Hutter, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-15}$, and zero weight decay. The semantic projector has a dedicated parameter group

whose learning rate follows a cosine decay without warmup over the first 30k optimization steps:

$$\eta_{\text{sem}}(s) = \begin{cases} \frac{\eta_0}{2}\left[1 + \cos\left(\frac{\pi s}{S}\right)\right], & 0 \le s < S, \\ 0, & s \ge S, \end{cases} \qquad S = 30{,}000, \tag{11}$$

where s is the optimization step within the 50k-step ablation run and η_0 is the initial learning rate of the semantic projector. This schedule is applied independently of the learning-rate schedule for the generative model. At step S, the projector learning rate reaches zero and remains zero thereafter; the associated normalization statistics are also frozen at this point. The projector parameters and normalization statistics remain fixed for the remaining 20k steps of the ablation run, while the generative model continues training. The learned projector and its normalization statistics are then reused without further updates throughout UniFacet-1.5B curriculum phases C1–C4 and SFT.

F.3 UniFacet-1.5B Training

We train UniFacet-1.5B using four multimodal curriculum phases, C1–C4, followed by supervised fine-tuning (SFT). Table 8 summarizes the optimization settings, input configurations, task sampling ratios, and parallelism for C1–C4 and SFT.

Training data. During curriculum training, text-to-image (T2I) and image-editing examples are drawn from our internal corpora containing 80M and 150M samples, respectively. Images for the image-to-text (I2T) task are sourced from COYO (Byeon et al., 2022), Open Images (Kuznetsova et al., 2020), and SA-1B (Kirillov et al., 2023). Visual understanding (UND) examples introduced in C4 are drawn from FineVision (Wiedmann et al., 2025). For SFT, we use the publicly released instruction-tuning data from BLIP3o-60k (Chen et al., 2025a) and ShareGPT-4o-Image (Chen et al., 2025b) for image generation and editing.

Optimization. Curriculum phases C1–C4 and SFT use AdamW (Loshchilov & Hutter, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-15}$, and zero weight decay. The learning rate is 1×10^{-4} for C1–C4 and 2×10^{-5} for SFT. We use 250 warmup steps at the start of each curriculum phase and SFT, followed by a constant learning rate. We use bfloat16 mixed precision, a gradient clipping threshold of 1.0, and an exponential moving average (EMA) of the trainable model parameters with a decay rate of 0.9999.

Curriculum and input configuration. C1 and C2 sample T2I and I2T examples at a 4:1 ratio for 50k and 100k optimization steps, respectively. The longest side of images used for generation is capped at 256 pixels in C1 and 512 pixels in C2–C4 and SFT. C3 introduces image editing and trains for 130k steps with a T2I:Edit:I2T sampling ratio of 2:2:1. C4 adds visual understanding and trains for 150k steps with a T2I:Edit:I2T:UND ratio of 3:4:0.5:2.5. Finally, SFT runs for 10k steps with a 1:1 T2I-to-editing ratio. C1–C4 use a maximum sequence length of 4,596 tokens, with an average of approximately 4,096 tokens.

Trainable modules and objectives. The pretrained visual backbone and UniFacet tokenizer remain frozen throughout C1–C4 and SFT, while the Mixture-of-Transformer-Experts layers are optimized for the corresponding task mixture. We optimize autoregressive next-token prediction for text outputs and continuous velocity prediction for image generation and editing, following the combined objective in Eq. 10. For continuous visual prediction, we use a timestep shift of 4.0 and a latent patch size of 1.

Distributed training and checkpointing. Training in C1–C4 and SFT runs on 256 GPUs using FSDP2 with a sharding group size of 8 GPUs.

F.4 Ablation Training Details

All ablation experiments in Tables 4–6 are conducted on 96 GPUs. Unless otherwise specified below, the optimization hyperparameters and FSDP2 sharding group size follow the main training configuration in Table 8.

Table 8: **Training curriculum for UniFacet-1.5B.** Columns C1–C4 and SFT report the corresponding training configurations. Task ratios are relative sampling weights; "–" denotes an unsampled task.

Configuration	C1	C2	C3	C4	SFT
Optimization schedule					
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-15}$)				
Learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	2×10^{-5}
Gradient clipping	1.0	1.0	1.0	1.0	1.0
Weight decay	0.0	0.0	0.0	0.0	0.0
Schedule	Constant	Constant	Constant	Constant	Constant
Warmup steps	250	250	250	250	250
EMA decay	0.9999	0.9999	0.9999	0.9999	0.999
Precision	bfloat16	bfloat16	bfloat16	bfloat16	bfloat16
Training steps	50k	100k	130k	150k	10k
Resolution					
Avg. sequence length	4K	4K	4K	4K	4K
Avg. global batch size	2740	1115	855	855	760
Und. resolution	-	$\leq 512^2$	$\leq 512^2$	$\leq 512^2$	-
Gen. resolution	$\leq 256^2$	$\leq 512^2$	$\leq 512^2$	$\leq 512^2$	$\leq 512^2$
Data mixture					
Understanding	-	-	-	2.5	-
Image-to-text	1	1	1	0.5	-
Text-to-image	4	4	2	3	1
Image-editing	-	-	2	4	1
Parallelism					
FSDP2 sharding group size	8	8	8	8	8
Total GPUs	256	256	256	256	256

Semantic preservation (Table 4). We use Qwen2.5-1.5B as the language backbone, without adding QK normalization or MoT layers. To assess whether residual completion preserves pretrained semantics, we train each understanding model in two successive phases: 30k optimization steps on I2T data, followed by 30k steps on UND data. For each semantic anchor, the models using z_{vfm} and z_{uni} follow the same two-phase schedule, and each representation is used consistently during training and evaluation. All other training settings match those of the main experiments.

Complementary generative facets (Table 5). We use the data mixture and input configuration of curriculum phase C2 in Table 8, with the longest image side capped at 512 pixels, and a separate training budget of 50k optimization steps for each variant. T2I and I2T examples are sampled at a $4{:}1$ ratio. For variants with a semantic facet, the semantic projector is jointly optimized with the generative model during the first 30k of the 50k optimization steps, following the schedule in Appendix F.2; the pretrained reconstruction pathway remains frozen throughout. The projector and its normalization statistics are fixed for the remaining 20k steps, while the generative model continues training. The reconstruction-only baselines and the variants with different semantic-facet dimensions use the same training budget.

Editing tokenizer configurations (Table 6). We first train for 150k optimization steps using the data mixture and input configuration of curriculum phase C2 in Table 8, then continue for another 150k steps using those of C3, for a total of 300k steps. The longest image side is capped at 512 pixels in both phases. The first phase samples T2I and I2T at a $4{:}1$ ratio; the second introduces editing with a T2I:Edit:I2T ratio of $2{:}2{:}1$. UniFacet and both reproduced BAGEL configurations follow this same schedule. The representation-swap variant marked with † is evaluated using the trained UniFacet model and requires no additional training.

Table 9: **Cross-representation evaluation.** Δ is measured relative to evaluation with the same representation used during training.

Anchor	Train	Inference	SEED	MMStar	MMB	OCR	RWQA	AI2D	Avg.↑	Δ Avg.
RADIOv4	z_{uni}	z_{vfm}	76.2	55.5	74.4	64.1	61.4	77.5	68.18	-0.10
RADIOv4	z_{vfm}	z_{uni}	76.2	54.9	73.6	63.7	61.4	78.3	68.01	$+0.18$
SigLIP2	z_{uni}	z_{vfm}	76.0	55.6	73.2	53.9	62.8	78.6	66.68	0.00
SigLIP2	z_{vfm}	z_{uni}	75.9	55.1	74.4	54.1	61.8	78.9	66.70	$+0.12$

Table 10: **Full comparison of image reconstruction quality, including both semantic anchors.** All methods are evaluated at 512×512 resolution on COCO and ImageNet, and additionally at 256×256 resolution on ImageNet. Higher PSNR and SSIM indicate better reconstruction quality. Within each group, the best result is shown in **bold** and the second best is underlined.

Method	Downsample	Channels	COCO 512		ImageNet 512		ImageNet 256	
			PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Reconstruction-oriented autoencoders								
Qwen-VAE (Wu et al., 2025a)	8×	16	33.60	0.92	32.96	0.92	30.19	0.88
FLUX-VAE (Black Forest Labs, 2024)	8×	16	**34.20**	**0.95**	**33.91**	**0.95**	**30.95**	**0.91**
NextStep-VAE (Team et al., 2025a)	8×	16	33.12	0.93	32.65	0.92	30.07	0.89
Hunyuan-3.0-VAE (Team, 2025)	16×	32	32.32	0.90	31.74	0.89	29.44	0.86
Unified visual tokenizers								
DualToken (Song et al., 2025)	14×	1152	27.06	0.81	26.70	0.79	26.07	0.80
DualViTok (Huang et al., 2025)	16×	64	24.09	0.71	23.82	0.69	21.54	0.61
UniTok (Ma et al., 2025)	16×	64	26.60	0.80	26.31	0.79	24.31	0.73
UniLIP (Tang et al., 2025)	32×	32	25.40	0.74	25.09	0.71	22.97	0.65
UniFacet-SigLIP2 (Ours)	16×	64	33.30	0.93	32.88	0.93	30.13	0.89
UniFacet-RADIOv4 (Ours)	16×	64	32.95	0.93	32.61	0.93	29.91	0.90
UniFacet-RADIOv4 (Ours)	16×	128	**35.95**	**0.97**	**35.77**	**0.97**	**32.60**	**0.95**

G Cross-Representation Compatibility

We further test whether the completed representation z_{uni} remains compatible with the semantic space of z_{vfm}. For each semantic anchor, we train the understanding model with one representation and replace it with the other only at inference, without retraining the model. Table 9 shows that this swap changes the average score by at most 0.18 points. This result provides additional evidence that residual completion preserves the semantic coordinate system of the pretrained VFM.

H Additional Reconstruction and Alignment Results

H.1 Full Quantitative Comparison

Table 10 expands the main-text comparison to include UniFacet-SigLIP2 alongside both UniFacet-RADIOv4 configurations, using the evaluation protocol in Appendix E.2. At 64 channels, the SigLIP2 variant achieves higher PSNR than the RADIOv4 variant in all three settings. The two variants have the same reported SSIM on COCO and ImageNet at 512×512, while RADIOv4 achieves higher SSIM on ImageNet at 256×256.



Figure 4: **Qualitative image reconstruction comparison.** From left to right: input images, UniFacet-RADIOv4 with a 64-channel reconstruction facet, FLUX VAE, DualToken, and UniLIP. Red and blue boxes mark corresponding regions, with enlarged crops shown directly below each image.

H.2 Qualitative Reconstruction Comparisons

Figure 4 presents side-by-side reconstruction results for four examples. The enlarged regions highlight differences in text legibility, facial details, and small scene structures.

Table 11: **Representation alignment after residual completion.** Mean per-image CKNNA between z_{vfm} and z_{uni} on 500 COCO 2017 validation images at 512×512 resolution, using $k = 10$. Channel counts refer to the reconstruction facet, rather than the representations used to compute CKNNA.

Tokenizer configuration	CKNNA
UniFacet-SigLIP2	0.995
UniFacet-RADIOv4 (64 channels)	0.996
UniFacet-RADIOv4 (128 channels)	0.995

H.3 CKNNA Alignment

We evaluate CKNNA between the frozen VFM representation z_{vfm} and the completed unified representation z_{uni} on 500 images from the COCO 2017 validation set at 512×512 resolution. The subset consists of the first 500 images after sorting filenames. For each image, we extract spatial tokens from both representations without global pooling and ℓ_2-normalize each token along the feature dimension. We compute CKNNA with $k = 10$ using inner-product similarities, excluding self-neighbors and using the unbiased HSIC estimator. Feature extraction is performed in evaluation mode with float32 precision, and the final score is the arithmetic mean of the per-image CKNNA scores.

Table 11 reports the results for the SigLIP2 anchor and the RADIOv4 anchor with 64- and 128-channel reconstruction facets.

All three configurations achieve CKNNA scores close to one, indicating strong alignment between the pretrained and completed representations. Together with the visual understanding and cross-representation evaluations in Section 4.4, these results support the preservation of pretrained semantic structure after residual completion.

I Direct Semantic Alignment of the Reconstruction Latent

We investigate whether directly imposing semantic alignment on the reconstruction latent provides a better reconstruction–generation trade-off than the complementary-facet design of UniFacet. Inspired by PS-VAE (Zhang et al., 2025a), which combines reconstruction of pretrained semantic features with pixel reconstruction through a compact latent, we evaluate a variant that applies semantic supervision directly to z_{rec}. This variant places semantic and reconstruction requirements on the same compact representation. UniFacet instead preserves the learned reconstruction pathway and adapts a separate semantic facet z_{sem} for joint generative modeling.

For reconstruction, we compare the direct-alignment variant with the default UniFacet-RADIOv4 tokenizer with $d_{\mathrm{rec}} = 64$; for generation, we use the $d_{\mathrm{sem}} = 32$ complementary-facet ablation variant from Table 5. Both generation experiments use the same 50k-step training configuration described in Appendix F.4. The direct-alignment results in Table 12 are obtained within our framework using the semantic-alignment strategy inspired by PS-VAE.

Table 12: **Direct semantic alignment versus complementary facets.** Reconstruction is evaluated using PSNR and SSIM, and generation is evaluated using GenEval under the same 50k-step training configuration. UniFacet reconstruction values retain the reported precision of Table 10; its GenEval score is from the $d_{\mathrm{sem}} = 32$ complementary-facet ablation in Table 5. All metrics are higher-is-better.

Method	COCO 512		ImageNet 512		ImageNet 256		GenEval
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	
Direct alignment	32.6167	0.9260	32.2891	0.9226	29.7451	0.8877	0.68
UniFacet	32.95	0.93	32.61	0.93	29.91	0.90	0.71

Direct alignment yields lower PSNR in all three reconstruction settings, with gaps of approximately 0.33, 0.32, and 0.16 dB on COCO at 512×512, ImageNet at 512×512, and ImageNet at 256×256, respectively. Its GenEval score is also lower, reaching 0.68 compared with 0.71 for UniFacet under the same training configuration. Thus, directly aligning the reconstruction latent underperforms our complementary-facet design in both reconstruction and generation in this experiment. These results support preserving the reconstruction role of z_{rec} while exposing generative semantic guidance through a separately adapted facet derived from the same unified visual space.

J Additional Text-to-Image Generation Results

Figures 5 and 6 present additional text-to-image generation results of UniFacet-1.5B across diverse subjects and visual styles.

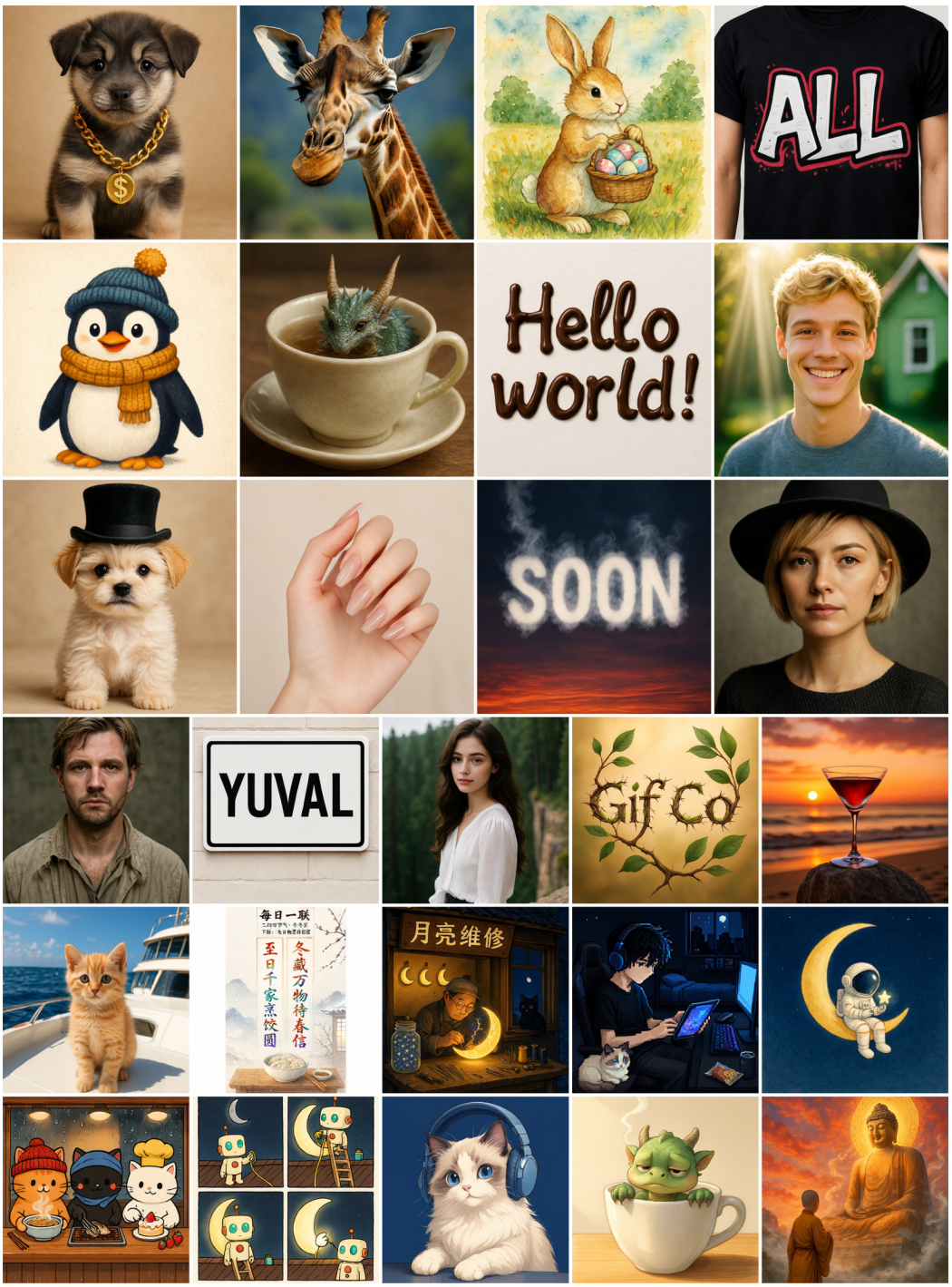

Figure 5: **Additional text-to-image generation results of UniFacet-1.5B (I).**

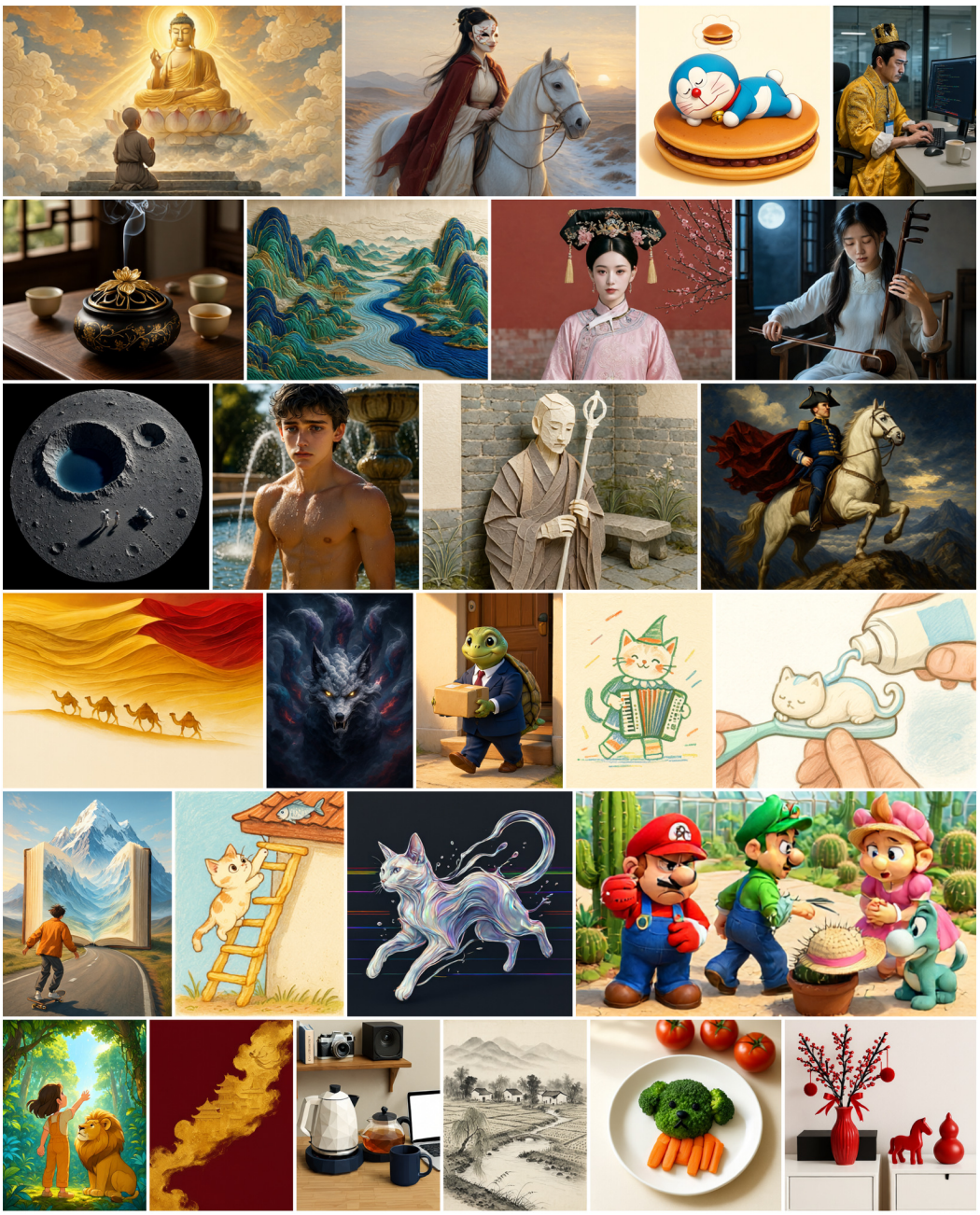

Figure 6: **Additional text-to-image generation results of UniFacet-1.5B (II).**

K FULL IMGEDIT RESULTS

Table 13 provides the full ImgEdit breakdown for the models compared in Table 3, including all nine editing categories and the overall score. The main table presents Add, Remove, Replace, Style, and Overall for a compact comparison.

Table 13: **Full quantitative comparison on ImgEdit.** Models are sorted by ascending ImgEdit overall score within each group. Baseline values follow the evaluation protocol reported by SenseNova-U1. Bg. denotes background editing, and Hybrid denotes composite editing (`compose` in the evaluation output). All scores are higher-is-better. **Bold** and underlined scores denote the best and second-best results across all listed models in each column, respectively, with ties marked equally. "–" denotes unreported values.

Model	# Params	Add	Adjust	Extract	Replace	Remove	Bg.	Style	Hybrid	Action	Overall↑
Closed-source Models											
FLUX.1-Kontext [Pro] (Labs et al., 2025)	–	4.25	4.15	2.35	4.56	3.57	4.26	4.57	3.68	4.63	4.00
GPT-Image-1 [High] (OpenAI, 2025b)	–	4.61	4.33	2.90	4.35	3.66	**4.57**	**4.93**	**3.96**	**4.89**	4.20
Nano-Banana (Google, 2025)	–	**4.62**	**4.41**	3.68	4.34	**4.39**	4.40	4.18	3.72	4.83	4.29
Open-source Models											
MagicBrush (Zhang et al., 2023)	0.9B	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.83
Instruct-Pix2Pix (Brooks et al., 2022)	0.9B	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88
AnyEdit (Yu et al., 2024)	1.3B	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
UltraEdit (Zhao et al., 2024)	2B	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.70
OmniGen (Xiao et al., 2024)	3.8B	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
ICEdit (Zhang et al., 2025b)	12B	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
Step1X-Edit (Liu et al., 2025)	12B	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
BAGEL (Deng et al., 2025)	7B	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
UniWorld-V1 (Lin et al., 2025)	12B	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
OmniGen2 (Wu et al., 2025b)	4B	3.57	3.06	1.77	3.74	3.20	3.57	4.81	2.52	4.68	3.44
InternVL-U (Tian et al., 2026)	1.7B	4.13	3.40	2.27	4.13	3.39	3.84	4.77	3.03	4.05	3.67
FLUX.1-Kontext [Dev] (Labs et al., 2025)	12B	4.12	3.80	2.04	4.22	3.09	3.97	4.51	3.35	4.25	3.71
Ovis-U1 (Wang et al., 2025)	1.2B	3.99	3.73	2.66	4.38	4.15	4.05	4.86	3.43	4.68	3.97
Qwen-Image-Edit (Wu et al., 2025a)	20B	4.38	4.16	3.43	**4.66**	4.14	4.38	4.81	3.82	4.69	4.27
Z-Image-Edit (Team et al., 2025b)	6B	4.40	4.14	**4.30**	4.57	4.13	4.14	4.85	3.63	4.50	**4.30**
UniFacet-1.5B (Ours)	1.5B	4.21	4.08	3.64	4.31	4.18	3.92	4.49	3.30	4.14	4.03

L Discussion of Perceptual Quality on GEdit-Bench

UniFacet-1.5B achieves strong semantic consistency on GEdit-Bench, while its perceptual quality (PQ) scores of 7.06 (EN) and 7.00 (CN) remain below those of the leading methods in PQ (Table 3). One potential contributor is our restricted output resolution: image editing uses a maximum long side of 512 pixels while preserving the image aspect ratio. In contrast, the official GEdit-Bench evaluation script resizes each image to approximately $512 \times 512 = 262{,}144$ total pixels while preserving its aspect ratio.[3] Thus, for non-square images, our native output contains fewer pixels than the evaluation grid.

Let $r = \max(W, H)/\min(W, H) \geq 1$ denote the ratio of the long side to the short side. For an output whose long side reaches 512 pixels, ignoring integer rounding of image dimensions, the native pixel count and its reduction relative to the evaluation grid are

$$N_{\text{native}} = \frac{512^2}{r}, \qquad \Delta N = 512^2 \left(1 - \frac{1}{r}\right), \qquad \frac{\Delta N}{N_{\text{eval}}} = 1 - \frac{1}{r}, \tag{12}$$

where $N_{\text{eval}} = 512^2$ is the target evaluation pixel count before integer rounding. For example, a 4:3 output has $512 \times 384 = 196{,}608$ native pixels, or 65,536 fewer pixels (25%); a 16:9 output has $512 \times 288 = 147{,}456$ native pixels, or 114,688 fewer pixels (43.75%). For a 3:2 aspect ratio, the reduction is approximately 33.3%; square images have no such reduction at a long side of 512 pixels.

The standard evaluation resizing upsamples these non-square outputs to the target pixel count, but cannot recover details absent from the native output. This can reduce apparent sharpness and fine detail, plausibly contributing to lower PQ, whose criteria assess naturalness and visual artifacts (Liu et al., 2025), with the official scoring prompt explicitly including blurred faces as an example of artifacts. The percentages above quantify the native pixel-count deficit, rather than a measured loss in PQ. We therefore regard resolution as a plausible contributing factor to the PQ gap; isolating its effect requires a controlled resolution ablation with all other settings fixed, which we leave for future work.

M Additional Image Editing Results

Figures 7 and 8 present additional image editing results of UniFacet-1.5B on the English and Chinese sets of GEdit-Bench (Liu et al., 2025), respectively. The examples cover text editing, object replacement and removal, background changes, attribute modification, and style transfer.

[3] Official GEdit-Bench evaluation script and PQ scoring prompt.



Figure 7: **Additional image editing results on GEdit-Bench with English instructions.** Each pair shows the source image on the left and the UniFacet-1.5B result on the right, with the editing instruction below.



Figure 8: **Additional image editing results on GEdit-Bench with Chinese instructions.** Each pair shows the source image on the left and the UniFacet-1.5B result on the right, with the editing instruction below.