UNISPACE: UNIFIED VISUAL REPRESENTATION AND SCALABLE MULTIMODAL MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic vision encoders provide a powerful interface for multimodal understanding, but their final tokens lose fine-grained visual details needed for reconstruction and image editing. We ask whether understanding, generation, and editing can share a visual representation built from a pretrained semantic ViT. A controlled probe shows that replacing only the patch embedding improves recoverability while keeping all Transformer blocks frozen. Motivated by this observation, we introduce *Patch Reparameterization*: the original semantic pathway is preserved, and a reconstruction-aware patch embedding supplies visual details to the same frozen backbone. We compress and concatenate the two token streams and balance their contributions during flow-matching training. The resulting tokenizers retain aggregate multimodal understanding while enabling high-fidelity reconstruction and image generation. PR-DINOv2 achieves an ImageNet reconstruction FID of 0.14 and a generation FID of 1.87 with classifier-free guidance. We further scale this representation into *UniSpace*, an 8B Mixture-of-Transformer-Experts model that supports understanding, generation, and editing without a separate VAE pathway. It scores 4.28 on ImgEdit and 0.84 on GenEval, demonstrating a shared ViT-based visual interface at system scale.

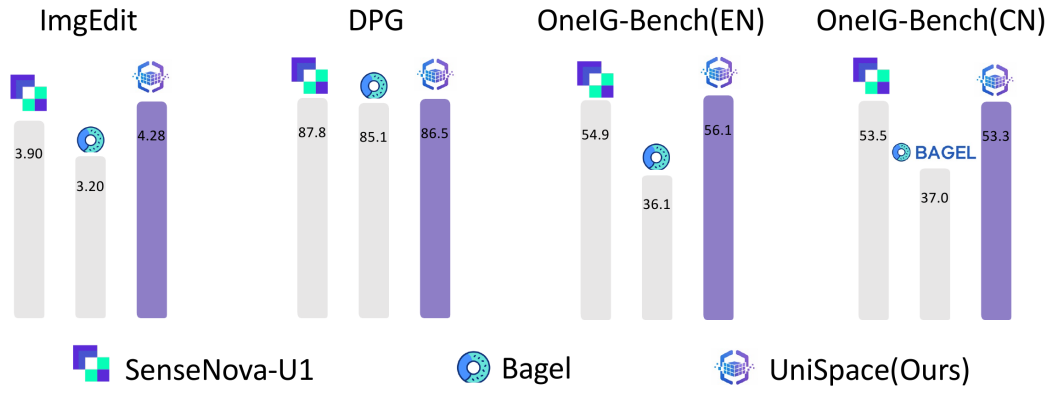

Figure 1: UniSpace generation and editing overview. Higher is better; OneIG scores are shown as percentages. Results use reported system settings, not matched compute; see Table 6.

1 INTRODUCTION

Semantic vision encoders such as CLIP (Radford et al., 2021) and SigLIP2 (Tschannen et al., 2025) support multimodal understanding and condition generation and editing. Yet their final tokens poorly preserve the fine-grained detail needed for reconstruction. In editing, a reference image must convey scene semantics alongside the colors, textures, and spatial structure that should survive the edit.

Current systems therefore often pair VAE latents for generation with semantic encoder features for understanding or guidance. Unified multimodal models such as BAGEL (Deng et al., 2025) bring these tasks into one architecture but retain the split between semantic features and generative latents.

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.

We ask whether a pretrained semantic ViT can instead provide one visual space for understanding, generation, and reference-preserving editing. Such a space must recover visual detail without losing the encoder's semantic prior, and its tokens must remain amenable to generative modeling at system scale.

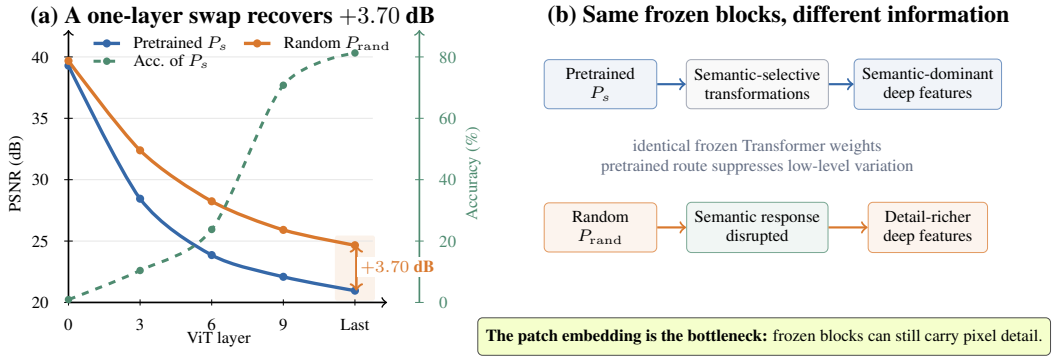

Figure 2: **The patch embedding—not the frozen Transformer blocks—is the reconstruction bottleneck.** (a) As semantic accuracy rises with depth, pixel recoverability through the pretrained SigLIP2 pathway collapses. Replacing only its patch embedding P_s with a random projection P_{rand} raises last-layer PSNR by 3.70 dB; all Transformer weights remain frozen and the reconstruction probes are identical. (b) This controlled intervention shows that input parameterization determines which information route the same backbone activates: the pretrained route favors semantic abstraction, whereas disrupting it exposes a detail-richer residual pathway.

We first test why semantic ViTs reconstruct poorly. In the controlled probe of Fig. 2, replacing only the patch embedding improves deep-layer pixel recoverability while all Transformer blocks remain frozen. The pretrained input pathway, rather than an intrinsic limit of those blocks, is the reconstruction bottleneck. This finding motivates *Patch Reparameterization*: we keep the original patch embedding for semantic understanding and learn a second, reconstruction-aware embedding through the same frozen backbone. We compress and concatenate the two token streams so that semantic and detail-bearing components remain explicit, and balance their flow-matching losses during generation. PR-SigLIP2, PR-DINOv2, and PR-Qwen-ViT instantiate this design. They preserve semantic capability while supporting high-fidelity reconstruction and image generation.

We then scale PR-Qwen-ViT into *UniSpace*, an 8B Qwen-based Mixture-of-Transformer-Experts model. Its patch-reparameterized ViT encoder–decoder is the sole visual interface: reference images, target images, and generated outputs share the same representation. UniSpace supports understanding, text-to-image generation, and instruction-based editing without a separate VAE space.

Our contributions are (1) a controlled diagnosis of the patch embedding as the detail bottleneck in frozen semantic ViTs; (2) a factorized visual representation and balanced generative objective, instantiated across three encoders; and (3) UniSpace, which applies that representation to understanding, generation, and editing in one multimodal system.

2 Related Work

2.1 Semantic Representations for Generation and Reconstruction

REPA (Yu et al., 2024) aligns diffusion-transformer features with pretrained semantic representations to improve convergence and generation. VA-VAE (Yao et al., 2025b) introduces semantic alignment into autoencoder training, while RAE (Zheng et al., 2025) pairs frozen semantic features with a learned decoder to use them directly as generative latents.

Unified visual tokenizers must also balance understanding, reconstruction, and generative modeling. UniFlow (Yue et al., 2025) distills understanding-aware reconstruction representations; RAEv2 (Singh et al., 2026) incorporates multi-layer DINO (Caron et al., 2021) features. Strong understanding and reconstruction do not necessarily yield latents that a generative prior can easily

model. Patch Reparameterization addresses this trade-off through reconstruction-aware inputs to a frozen semantic ViT.

2.2 Visual Representation Spaces in Unified Multimodal Models

Unified multimodal models differ in their visual interfaces (Cui et al., 2025; Wang et al., 2024; SenseNova-U1 Team, 2026; Deng et al., 2025). Emu3-family models use discrete visual tokens (Cui et al., 2025; Wang et al., 2024). BAGEL (Deng et al., 2025) uses semantic encoder tokens for understanding and VAE latents for generation, connected through shared self-attention in a Mixture-of-Transformer-Experts architecture. Its architecture is unified, but its visual spaces remain separate.

UniSpace instead uses a patch-reparameterized ViT encoder–decoder as its sole visual interface, preserving the pretrained semantic prior while supporting understanding, generation, and editing without a separate VAE latent space.

3 Unified Encoder via Patch Reparameterization

3.1 Patch Reparameterization Can Carry Visual Details

Semantic ViTs retain high-dimensional tokens throughout their Transformer blocks, yet deeper features lose pixel-level recoverability as semantic accuracy increases. Is this loss intrinsic to the frozen blocks, or does it depend on their input parameterization? We test this on SigLIP2 by replacing only the pretrained patch embedding with a random linear projection, keeping all Transformer blocks frozen and training identical reconstruction probes at each depth (Fig. 2).

The pretrained and random projections have similar patch-level PSNR (39.29 and 39.68), but last-layer PSNR improves from 20.96 to 24.66 under the random projection. The same frozen blocks can therefore carry more recoverable visual detail under a different input parameterization. This is consistent with the pretrained pathway selectively suppressing detail as it abstracts semantics, rather than an unavoidable dimensional bottleneck. Randomization also disrupts the pretrained semantic representation; we instead preserve that pathway and learn a second patch embedding for reconstruction.

3.2 Constructing a Unified Representation

Let P_s be the pretrained patch embedding and F_ϕ the frozen Transformer blocks. A second patch embedding P_r, initialized from P_s, is trained for reconstruction. The two pathways produce

$$T_s = F_\phi(P_s(I)), \qquad T_r = F_\phi(P_r(I)). \tag{1}$$

Both pathways share frozen Transformer parameters but require separate feature evaluations. The original P_s also stays frozen, so the semantic tokens are structurally unchanged throughout reconstruction training.

Why not merge the features implicitly? An auxiliary diagnostic merges semantic and reconstruction features with an MLP into an entangled latent $Z_m = M(Z_s, Z_r)$, aligned to both source spaces. It preserves zero-shot accuracy (78.53 versus 79.10 for SigLIP) and reconstructs real encoded images well (PSNR 33.83, rFID 0.069). Yet a DiT prior's generated latents yield FID 120.9 with the high-fidelity reconstruction decoder, versus 8.07 with the semantic decoder. Approximately 95% of merger output variation is explained by the semantic pathway. This decoder-dependent gap suggests that the prior captures semantic structure while missing reconstruction-critical directions: good understanding and reconstruction alone do not ensure generative modelability. The diagnostic is separate from our final tokenizer; its full protocol and visualizations are in Appendix F.

Explicit Token Fusion. We retain separately addressable channel groups instead of an entangled merger. A learnable linear projection compresses reconstruction features, then concatenation preserves their boundary with the semantic features:

$$\widetilde{T}_r = W_r T_r, \qquad T_u = \mathrm{Concat}(T_s, \widetilde{T}_r). \tag{2}$$

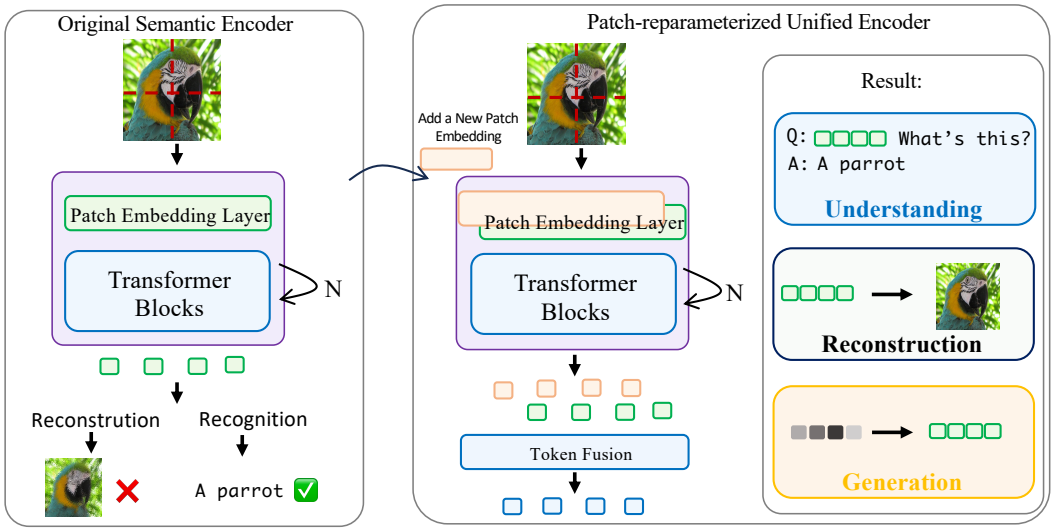

Figure 3: Overview of Patch Reparameterization. The original semantic patch embedding and frozen ViT blocks preserve the pretrained semantic pathway, while a reconstruction-aware patch embedding injects visual details into the same frozen backbone. The semantic and reconstruction-aware tokens are combined by a Token Fusion layer to form the unified representation, which is shared by understanding, reconstruction, and generation.

Compression lowers the generative model's dimensional burden; concatenation makes each component accessible to its own training weight. Neither operation changes T_s. The complete T_u is used for understanding, reconstruction, and generation, not either component alone. We use $d_r = 128$, giving $768+128 = 896$ channels for PR-SigLIP2 and PR-DINOv2, and $1152+128 = 1280$ for PR-Qwen-ViT. This choice trades some reconstruction capacity for generative modelability (Table 5).

3.3 Training for Reconstruction and Generation

Reconstruction training. A ViT decoder D_ψ reconstructs the input with pixel and perceptual losses:

$$\widehat{I} = D_\psi(T_u), \qquad \mathcal{L}_{\text{rec}} = \mathcal{D}(\widehat{I}, I). \tag{3}$$

Only P_r, W_r, and D_ψ are updated; gradients through the frozen F_ϕ train the new input projection while P_s remains fixed. A subsequent decoder-only stage adds an adversarial loss. SigLIP2 and DINOv2 use ImageNet-1K, while Qwen-ViT uses web data. The ImageNet generation experiments additionally use decoder calibration on latents sampled from an early DiT checkpoint, keeping the encoder and generative model fixed. Appendix A provides architecture and optimization details.

Balanced flow matching. The explicit decomposition lets us control how much the generative objective focuses on decoder-critical reconstruction information.

Let $Z_1 = T_u$ and $Z_0 \sim \mathcal{N}(0, \mathbf{I})$. For time $t \in [0, 1]$, we use the linear flow path and its target velocity,

$$Z_t = (1-t)Z_0 + tZ_1, \qquad V_t = Z_1 - Z_0, \qquad \widehat{V}_t = v_\theta(Z_t, t, c), \tag{4}$$

where c is the task condition. We partition the target and predicted velocities into semantic and reconstruction components, denoted by superscripts s and r. For N tokens with a semantic and b reconstruction channels, our objective is

$$\mathcal{L}_{\text{BFM}} = \mathbb{E}\left[(1-\lambda_r)\frac{\|\widehat{V}_t^s - V_t^s\|_F^2}{Na} + \lambda_r \frac{\|\widehat{V}_t^r - V_t^r\|_F^2}{Nb}\right]. \tag{5}$$

Unlike a plain MSE over all channels, this loss separates dimensionality from the assigned objective weight. We set $\lambda_r = 0.75$, supported by the controlled ablation in Table 4. Generative training

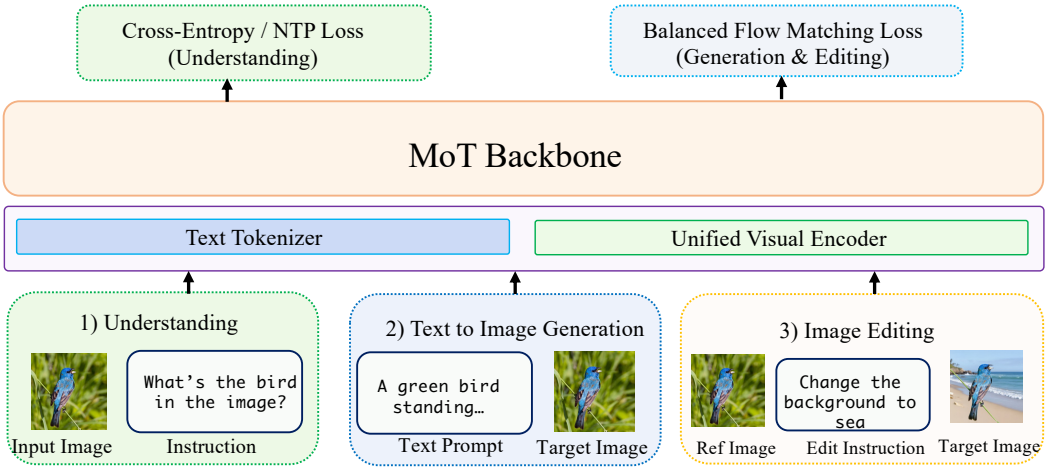

Figure 4: UniSpace pipeline built on the proposed unified visual representation. UniSpace uses the PR-Qwen-ViT encoder–decoder as the only visual interface for reference images, target images, and generated images. Text and visual tokens are routed through a Mixture-of-Transformer-Experts backbone for understanding, text-to-image generation, and instruction-based image editing, without an additional VAE pathway.

uses logit-normal timestep sampling and the dimension-dependent time shift of RAE (Zheng et al., 2025). At inference, integrating $\mathrm{d}Z_t/\mathrm{d}t = v_\theta(Z_t, t, c)$ from Gaussian noise gives $\widehat{T}_u$, which the shared decoder maps to an image.

For $768{+}128$ channels, a plain mean squared error assigns the reconstruction component only $1/7$ of the normalized loss weight. Our component-wise objective decouples this coefficient from channel count, allowing compression without automatically reducing reconstruction's objective weight. The flow-weight and channel-count ablations test these choices separately.

4 UniSpace: Scaling Unified Multimodal Modeling

Architecture. UniSpace builds on Qwen3-8B (Yang et al., 2025) and adopts the Mixture-of-Transformer-Experts design of BAGEL (Deng et al., 2025). Text and conditioning-image tokens are routed to the understanding expert; noised target-image tokens are routed to the generation expert. Shared self-attention permits interaction across the two experts. Unlike a dual ViT–VAE system, all reference and target images use PR-Qwen-ViT's T_u space, and generated latents use the same reconstruction decoder.

Tasks and objectives. For understanding, T_u and the instruction condition autoregressive textual responses, supervised by next-token cross-entropy. For text-to-image generation, text conditions balanced flow matching in T_u. Editing adds clean reference-image tokens to the instruction while predicting the noised target in that same space. Thus, generation and editing share their visual target and loss, differing only in conditioning. The tokenizer and decoder remain frozen during all UniSpace training; the MoT backbone and multimodal projections are trainable. Block-wise packing prevents attention across independent examples, with causal text attention and full visual attention within each example.

For a response $y_{1:L}$, the language loss is $\mathcal{L}_{\mathrm{NTP}} = -\sum_{k=1}^{L} \log p_\theta(y_k \mid y_{<k}, c_{\mathrm{und}})$. Task type selects the text or visual objective; the curriculum sets their frequencies. Table 8 compares visual interfaces.

Training scale. We train on internally curated generation, editing, and understanding data, progressing through 256, 512, and 1024 image resolutions, followed by multimodal supervised fine-tuning. Generation and understanding are mixed 10:1 in the first stage; later pretraining uses generation, editing, and understanding at 10:3:1, and SFT uses 10:3:2. The model sees approximately

510M sample instances and 470B multimodal tokens; sample instances include repeated draws and are not unique dataset size. The selected training trajectory uses approximately 142K Ascend 910B NPU-hours. Appendix B records the curriculum, hardware, optimization, and packing details.

5 EXPERIMENTS

5.1 UNIFIED TOKENIZER EVALUATION

The tokenizer must recover visual detail without losing the pretrained semantic prior, and its tokens must remain usable for generation. We test these requirements in turn. Reconstruction uses ImageNet-1K validation images at 256×256 with PSNR, SSIM, and reconstruction FID (rFID). For understanding, we compare each encoder with its original counterpart under the same LLaVA-v1.5 (Liu et al., 2024a) setup: frozen encoder, Vicuna-7B-v1.5, identical training data, and a two-layer MLP fed the complete T_u. For generation, a DiT prior follows RAE (Zheng et al., 2025), with 50-step Euler sampling. Further protocols and results are in Appendix C.

Table 1: ImageNet reconstruction. RAEv2 uses DINOv3-L ($K = 7$); full settings: Table 10.

Tokenizer	PSNR↑	SSIM↑	rFID↓
RAE (SigLIP2-B)	19.35	0.49	0.53
RAE (DINOv2-B)	18.86	0.48	0.57
UniFlow (DINOv2)	31.01	0.94	0.54
RAEv2 (DINOv3-L)	22.57	0.63	0.29
PR-SigLIP2	29.64	0.87	0.18
PR-DINOv2	30.84	0.90	0.14
PR-Qwen-ViT	30.16	0.88	0.17

Reconstruction. Against RAE with the corresponding pretrained encoders, Patch Reparameterization improves all three reconstruction metrics (Table 1): PR-DINOv2 cuts rFID from 0.57 to 0.14 and raises PSNR from 18.86 to 30.84. Against UniFlow-DINOv2, its rFID is lower but its PSNR and SSIM are also lower. The benefit depends on the metric and comparator. Figure 5 illustrates clearer local character structure than RAE, though fine text remains imperfect.

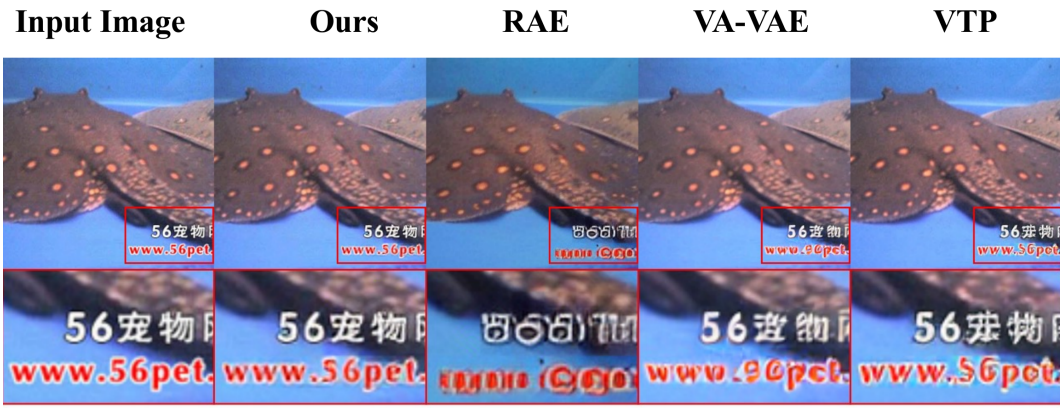

Figure 5: Reconstruction and detail enlargements: input, ours, RAE, VA-VAE, and VTP. Full comparison: Fig. 8.

Table 2: Understanding. Published rows are not matched; PR rows use full T_u. Encoder, LLM, resolution, and Avg. details: Table 11.

Method	POPE	GQA	TQA	MMV	MMB	MME-S	MME-P	Avg.
UniFlow-LV (Yue et al., 2025)	88.97	63.35	61.85	36.60	67.10	1803.0	1505.1	69.04
TokenFlow-XL (Qu et al., 2025)	87.80	62.50	62.30	48.20	76.80	1922.2	1551.1	73.04
SigLIP2-B	85.20	61.40	54.36	25.00	64.40	1689.3	1378.2	63.39
PR-SigLIP2	84.80	61.10	54.60	25.50	65.20	1741.1	1447.2	64.37
Qwen-ViT	86.26	63.25	64.27	31.88	69.24	1791.3	1471.3	68.29
PR-Qwen-ViT	86.57	63.13	63.68	30.28	69.07	1874.5	1522.7	68.94

Understanding. Better reconstruction would be of limited use if it erased the pretrained semantic prior. With full T_u, average scores rise from 63.39 to 64.37 for SigLIP2 and 68.29 to 68.94 for Qwen-ViT (Table 2), although some individual scores fall. UniFlow-LV and TokenFlow-XL provide recent context, but their encoders, resolutions, and LLMs differ from ours. The matched results support preserved aggregate understanding, not uniform gains or stronger full-system VLM performance.

Table 3: ImageNet reconstruction–generation trade-off (lower is better). CFG: classifier-free guidance. Reported priors and budgets differ across baselines; full metrics: Table 12.

Tokenizer / prior	rFID	gFID w/o CFG	gFID w/ CFG
VA-VAE / LightningDiT	0.28	2.17	1.35
VTP-L / LightningDiT	0.36	1.85	1.11
RAE-DINOv2 / DiT	0.57	1.51	1.13
RAEv2-DINOv3 / DiT	0.29	1.65	1.06
PR-SigLIP2 / DiT	0.18	4.42	2.80
PR-DINOv2 / DiT	0.14	2.10	1.87

Generation. Good reconstruction and understanding do not guarantee modelable latents. PR-DINOv2 reaches gFID 2.10 without guidance and 1.87 with CFG scale 1.2, alongside rFID 0.14 (Table 3). RAE, RAEv2, and VTP-L achieve lower gFID but higher rFID. Together with the understanding results, this supports a three-task balance, with room to improve generation.

5.2 Ablations

We test how the reconstruction component's flow-matching weight and channel dimension affect generation.

Table 4: Flow-weight ablation with a fixed 768 + 128 representation and training protocol.

λ_r	0.25	0.60	0.75	0.82	0.86	0.92
FID↓	8.99	7.42	**7.10**	7.24	7.13	7.39
IS↑	142.4	153.2	**154.4**	152.8	152.5	150.0

Balancing the generative objective. With the representation and training settings fixed, increasing the reconstruction weight from $\lambda_r = 0.25$ to 0.75 lowers FID from 8.99 to 7.10; at 0.92, FID rises to 7.39 (Table 4). Detail-bearing channels thus benefit from substantial, but not exclusive, weight. The full sweep appears in Table 19.

Table 5: Reconstruction-channel ablation with the same semantic stream and early generation budgets; full results: Table 20.

d_r	PSNR↑	rFID↓	FID@20↓	FID@40↓
128	29.79	0.163	**9.95**	**6.92**
768	**33.61**	**0.085**	15.51	11.07

Compressing reconstruction features. More reconstruction channels improve the encoded images but make the prior harder to train under the tested budgets. Increasing d_r from 128 to 768 lowers rFID from 0.163 to 0.085 but raises generation FID from 6.92 to 11.07 after 40 epochs (Table 5); the same trend holds at 20 epochs. This motivates the compact stream, while the fully trained generation comparison remains a separate experiment.

5.3 System-Level Generation and Editing

We next test the shared space at UniSpace scale. Editing requires the model to read a reference, follow an instruction, and preserve unrelated details; text-to-image generation uses the same output space without a reference. Training data and architecture also affect these system comparisons.

Table 6: Selected unified-model results; higher is better. OneIG is English/Chinese. Baselines follow SenseNova-U1 (SenseNova-U1 Team, 2026); full tables: Appendix C. Parameters do not imply matched compute.

Model	Params	ImgEdit	GenEval	OneIG EN/CN	DPG
BAGEL	7B	3.20	0.82	0.361/0.370	85.07
SenseNova-U1	8B	3.90	0.91	0.549/0.535	87.78
Emu3.5	32B	4.41	0.73	0.564/0.529	–
UniSpace	8B	4.28	0.84	0.561/0.533	86.49

Image editing. On the nine-category ImgEdit benchmark (Ye et al., 2025), UniSpace scores 4.28, above BAGEL and SenseNova-U1 but below the larger Emu3.5 (Table 6). On English/Chinese GEdit it scores 7.41/7.38, with strong semantic consistency but weaker perceptual quality on sensitive edits. Because GPT-4o judged our GEdit outputs, judge and protocol differences limit comparison with reported baselines. Full results and examples appear in Appendix C.

Table 7: Selected ImgEdit categories; higher is better. Background (Bg.) and hybrid edits complement local object operations. All nine categories and additional models are reported in Table 13.

Model	Add	Replace	Remove	Bg.	Style	Hybrid	Overall
BAGEL	3.56	3.30	2.62	3.24	4.49	2.38	3.20
SenseNova-U1	3.83	4.32	3.26	4.18	4.85	3.03	3.90
Emu3.5	4.61	4.84	4.58	4.35	4.79	3.69	4.41
UniSpace	4.53	4.67	4.42	4.23	4.55	2.70	4.28

The category breakdown shows UniSpace close to Emu3.5 on addition, replacement, and removal, with a larger gap on hybrid edits (Table 7). The object-addition example (Fig. 6) tests scene preservation as well as instruction following; outputs differ in object scale and placement.

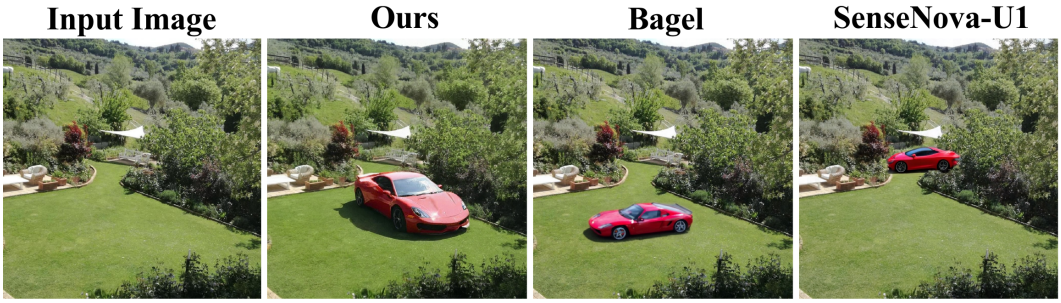

Figure 6: Object addition: input, UniSpace, BAGEL, and SenseNova-U1, with the original instruction. Additional edits: Figs. 9 and 10.

Text-to-image generation. Without a reference image, we measure object-centric composition with GenEval (Ghosh et al., 2023), English/Chinese generation with OneIG-Bench (Chang et al., 2025), and dense-prompt following with DPG-Bench (Hu et al., 2024). UniSpace scores 0.84, 0.561/0.533, and 86.49, respectively (Table 6). SenseNova-U1 is higher on GenEval and DPG, whereas UniSpace is higher on ImgEdit.

Category results locate the remaining weaknesses. UniSpace's DPG entity and relation scores (92.26 and 94.97) exceed its global-constraint score (84.80; Table 18). On Chinese OneIG, style exceeds SenseNova-U1 (0.455 versus 0.392), but text rendering falls below it (0.881 versus 0.977; Table 17). Counting and color binding on GenEval also remain difficult. The shared-prompt example (Fig. 7) shows recognizable subjects, with differences in pose, foliage, and lighting.

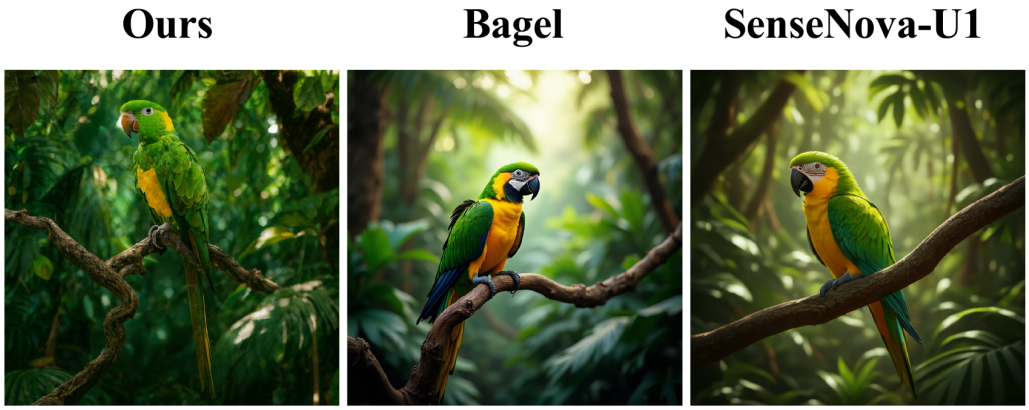

Figure 7: Text-to-image generation: UniSpace, BAGEL, and SenseNova-U1, with the original prompt excerpt. Full comparisons: Figs. 11 and 12.

6 Limitations and Conclusion

UniSpace prioritizes generation and editing; a frozen semantic pathway does not ensure gains on every VLM task, and optimizing it for understanding may hurt generation. The two visual paths add computation. Internal training data limit independent reproduction, and small benchmark differences lack significance estimates. Future work should test joint optimization, controlled system ablations, and broader understanding and safety. Patch Reparameterization nonetheless gives UniSpace one factorized visual space for understanding, generation, and editing without a separate VAE.

AI USE STATEMENT

Generative AI tools assisted with manuscript restructuring, language editing, LaTeX preparation, and submission-format checks. GPT-4o was used as an automatic judge for the UniSpace GEdit evaluation reported in Table 14. AI-assisted submission preparation did not execute the reported model-training experiments or replace their recorded measurements. The authors are responsible for the final text, claims, artifacts, and disclosure of AI use.

ETHICS STATEMENT

Image generation and editing can be misused to create deceptive content, alter a person's apparent identity or actions without consent, or reproduce biases and private information present in training data. This is particularly relevant to the human-centric editing examples in this work. Benchmark performance alone does not establish safety, fairness, or consent to use depicted individuals. The use of internally curated web data also raises questions of provenance, privacy, and usage rights that must be addressed before redistribution. Any deployment or release should include appropriate data-rights review, safeguards against non-consensual manipulation, and evaluation beyond the capability benchmarks reported here.

REPRODUCIBILITY STATEMENT

Section 3 specifies the unified representation and generative objective. Appendix A gives tokenizer architectures and training settings; Appendix B records the UniSpace curriculum, optimizer, hardware, and sequence-packing procedure. Appendix C contains complete comparison tables and qualitative results, while Appendices D, E, and F describe the ablations and diagnostic protocol. ImageNet-based tokenizer experiments use a public benchmark. The anonymous source-code supplement provides tokenizer and UniSpace inference and evaluation code plus release configurations; it does not include model checkpoints or internal training data. Those data limit fully independent reproduction of system-scale training.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

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

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

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*, 2025b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

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

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*, 2023.

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.

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

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

Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Feize Wu, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, et al. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 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*, 2025a.

Haokun Lin, Teng Wang, Yixiao Ge, Yuying Ge, Zhichao Lu, Ying Wei, Qingfu Zhang, Zhenan Sun, and Ying Shan. Toklip: Marry visual tokens to clip for multimodal comprehension and generation. *arXiv preprint arXiv:2505.05422*, 2025b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

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.

Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, et al. Tuna: Taming unified visual representations for native unified multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15740–15751, 2026.

Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024.

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. *Advances in Neural Information Processing Systems*, 38:129274–129297, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. In *Transactions on Machine Learning Research*, 2024.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2545–2555, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

SenseNova-U1 Team. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

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.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

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.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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.

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.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 93620–93638, 2025c.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems*, 38:47490–47518, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jingfeng Yao, Yuda Song, Yucong Zhou, and Xinggang Wang. Towards scalable pre-training of visual tokenizers for generation. *arXiv preprint arXiv:2512.13687*, 2025a.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025b.

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.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Zhengrong Yue, Haiyu Zhang, Xiangyu Zeng, Boyu Chen, Chenting Wang, Shaobin Zhuang, Lu Dong, Yi Wang, Limin Wang, and Yali Wang. Uniflow: A unified pixel flow tokenizer for visual understanding and generation. *arXiv preprint arXiv:2510.10575*, 2025.

Yue Zhao, Fuzhao Xue, Scott Reed, Linxi Fan, Yuke Zhu, Jan Kautz, Zhiding Yu, Philipp Krähenbühl, and De-An Huang. Qlip: Text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation. *arXiv preprint arXiv:2502.05178*, 2025.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

A Extended Tokenizer Method and Training Details

Pretrained semantic ViTs, such as SigLIP and CLIP, provide strong visual perception and image-text alignment, but their final tokens are not designed for faithful pixel reconstruction. Existing attempts to obtain understanding-aware reconstruction representations often train a new visual encoder (Yao et al., 2025a) or rely on semantic distillation (Yue et al., 2025), which complicates the system and can alter the original semantic representation. We ask a simpler question: is it necessary to relearn the Transformer blocks, or can a frozen semantic ViT be reparameterized to expose the visual details that its original semantic tokens suppress?

We answer this question with Patch Reparameterization. The pretrained patch embedding is kept unchanged to preserve the original semantic pathway, while an additional reconstruction-aware patch embedding provides an alternative parameterization that preserves reconstruction-relevant visual information within the same frozen Transformer blocks. Through a simple Token Fusion strategy, where reconstruction tokens are compressed and concatenated with semantic tokens, the resulting representation unifies semantic and reconstruction information, providing a shared visual interface for understanding and generative modeling.

A.1 Patch Reparameterization Can Carry Visual Details

Prior analyzes (Singh et al., 2026) have suggested that shallow features in semantic ViTs preserve more fine-grained visual details, whereas deeper features become more semantically abstract. Consistent with this observation, Fig. 2 shows that as visual tokens propagate through deeper Transformer blocks, their semantic classification accuracy increases from 0.93% to 81.29%, while their pixel-level recoverability degrades rapidly. However, this degradation does not necessarily imply that the frozen Transformer blocks are intrinsically unable to carry visual details. Unlike autoencoders with an explicit low-dimensional bottleneck, ViT backbones maintain high-dimensional token representations throughout the network. This design provides a plausible pathway for propagating input-dependent details across layers.

To test whether the poor recoverability is imposed by the frozen Transformer blocks themselves, we conduct a diagnostic experiment on SigLIP2. We replace its pretrained patch embedding with a randomly initialized linear projection, keep all subsequent Transformer blocks frozen, and train identical reconstruction probes on features extracted at different depths. If the frozen semantic Transformer blocks were incapable of carrying pixel-level information, changing only the input projection would not improve recoverability at deep layers.

Counterintuitively, the random patch embedding improves the last-layer PSNR from 20.96 to 24.66, despite using exactly the same frozen Transformer blocks. As shown in Fig. 2, at the patch-embedding output, the pretrained and random projections are similarly recoverable, achieving PSNR values of 39.29 and 39.68, respectively. Their behavior diverges only as the tokens pass through the same frozen semantic Transformer blocks. Under the pretrained patch embedding, the input activates visual patterns that the subsequent blocks were optimized to semantically abstract, progressively suppressing variations irrelevant to the semantic objective. The random projection disrupts this preferred semantic processing trajectory, allowing more low-level input variation to remain recoverable through the residual pathway.

This result suggests that pixel-level information is not lost because the hidden space cannot carry it. Rather, it is selectively suppressed under the semantic encoder's original input parameterization. However, randomizing the patch embedding is not itself a useful solution, since it destroys the pretrained semantic representation. It instead motivates a minimal adaptation: keep the original semantic patch embedding for understanding, and learn an additional reconstruction-aware patch embedding that injects visual details into the same frozen Transformer blocks, as shown in Fig.3.

A.2 Constructing a Unified Representation

Given an input image I, we construct a unified representation that preserves the semantic ability of the pretrained encoder while adding reconstruction details in a channel-factorized form. Let F_ϕ denote the frozen Transformer blocks of a pretrained semantic ViT, and let P_s denote its original pretrained patch embedding. The original semantic tokens are

$$T_s = F_\phi(P_s(I)), \tag{6}$$

where both P_s and F_ϕ are kept frozen. This path preserves the pretrained semantic representation and provides the basis for visual understanding.

To capture reconstruction-relevant details, we introduce a reconstruction-aware patch embedding P_r, initialized from P_s and optimized for reconstruction. It is followed by the same frozen Transformer blocks:

$$T_r = F_\phi(P_r(I)). \tag{7}$$

Thus, T_r is not produced by an independent reconstruction encoder; it is encoded within the same semantic ViT parameter space through a different input parameterization.

Explicit Token Fusion. A naive way to unify T_s and T_r is to fuse them into a single entangled feature space using an MLP. However, as discussed in Sec. C.1.4, this implicit fusion makes it difficult to control the relative contributions of semantic and reconstruction information during generative modeling. We therefore preserve the two components as explicitly separated feature groups along the channel dimension. Before concatenation, we first project the reconstruction tokens into a compact space:

$$\widetilde{T}_r = W_r T_r, \tag{8}$$

where W_r is a learnable linear projection along the channel dimension, mapping the reconstruction tokens from d to a lower dimension d_r. This projection serves a distinct role from concatenation: it reduces the dimensionality of the reconstruction component, lowering the burden for the generative model while retaining decoder-critical visual details. We then concatenate the original semantic tokens and the compressed reconstruction tokens along the channel dimension:

$$T_u = \mathrm{Concat}(T_s, \widetilde{T}_r). \tag{9}$$

The complete unified representation T_u serves as the shared visual representation for all downstream tasks, including understanding, reconstruction, and generation; neither T_s nor $\widetilde{T}_r$ is used independently. Unlike learned feature merging, this explicit fusion leaves the pretrained semantic representation unchanged and maintains an explicit boundary between semantic and reconstruction information. Compression reduces the dimensionality of reconstruction details, while concatenation makes the two information sources separately addressable during generative training.

A.3 Training for Reconstruction and Generation

Training the Reconstruction-Aware Path. After constructing T_u, we adopt a ViT decoder D_ψ, following RAE, to reconstruct the input image:

$$\hat{I} = D_\psi(T_u). \tag{10}$$

During training, the reconstruction-aware components and the decoder are optimized with a reconstruction objective:

$$\mathcal{L}_{\mathrm{rec}} = \mathcal{D}(\hat{I}, I), \tag{11}$$

where $\mathcal{D}$ is instantiated with pixel-level and perceptual reconstruction losses. Only the reconstruction-aware patch embedding P_r, the linear projection W_r, and the decoder D_ψ are updated. The original patch embedding P_s and the pretrained ViT blocks F_ϕ remain frozen. Since reconstruction training never modifies the original semantic path, the pretrained semantic tokens are structurally preserved while the additional path learns to supply decoder-critical details.

Balanced Generative Modeling. The explicit channel-wise decomposition of T_u also gives direct control over generative training. A plain MSE over the concatenated representation weights the two components according to their dimensionality and scale, without specifying how much of the objective should focus on decoder-critical reconstruction information. We therefore use a balanced flow-matching objective that normalizes each component by its dimension and then assigns an explicit objective weight.

We model the distribution of the unified representation using conditional flow matching. Let $Z_1 = T_u$ denote a unified representation, $Z_0 \sim \mathcal{N}(0, \mathbf{I})$ a Gaussian noise sample of the same shape, and $t \sim \mathcal{U}[0, 1]$ a timestep. We construct the linear probability path

$$Z_t = (1 - t)Z_0 + tZ_1, \tag{12}$$

whose target velocity is

$$V_t = \frac{\mathrm{d}Z_t}{\mathrm{d}t} = Z_1 - Z_0. \tag{13}$$

Given Z_t, t, and condition c, the generative model v_θ predicts the velocity $\widehat{V}_t = v_\theta(Z_t, t, c)$. Let $T_s \in \mathbb{R}^{N\times a}$ and $\widetilde{T}_r \in \mathbb{R}^{N\times b}$ denote the semantic and reconstruction components of T_u, respectively. Their explicit channel-wise separation allows us to partition the target and predicted velocities as

$$V_t = \mathrm{Concat}(V_t^s, V_t^r), \qquad \widehat{V}_t = \mathrm{Concat}(\widehat{V}_t^s, \widehat{V}_t^r). \tag{14}$$

We normalize the prediction error of each component by its number of dimensions and optimize the reconstruction–semantic balanced flow-matching objective

$$\mathcal{L}_{\mathrm{BFM}} = \mathbb{E}_{Z_1, Z_0, t, c}\left[(1 - \lambda_r)\frac{\|\widehat{V}_t^s - V_t^s\|_F^2}{Na} + \lambda_r \frac{\|\widehat{V}_t^r - V_t^r\|_F^2}{Nb}\right], \tag{15}$$

where λ_r specifies the fraction of the total objective weight assigned to the reconstruction component. We set $\lambda_r = 0.75$, allocating 75% of the objective weight to the decoder-critical reconstruction component while retaining the semantic component that facilitates generative learning; this choice is supported by the ablation in Table 19.

At inference time, we sample $Z_0 \sim \mathcal{N}(0, \mathbf{I})$ and solve the ordinary differential equation

$$\frac{\mathrm{d}Z_t}{\mathrm{d}t} = v_\theta(Z_t, t, c), \qquad t : 0 \to 1, \tag{16}$$

to obtain a generated unified representation $\widehat{T}_u$, which is decoded into an image as $\widehat{I} = D_\psi(\widehat{T}_u)$.

A.4 Implementation Details

Unified tokenizer construction. We instantiate Patch Reparameterization on three pretrained semantic ViTs: SigLIP2-B (Tschannen et al., 2025), DINOv2-B (Oquab et al., 2024), and Qwen-ViT (Bai et al., 2025), obtaining PR-SigLIP2, PR-DINOv2, and PR-Qwen-ViT, respectively. SigLIP2-B is evaluated at 256×256 resolution with a 16×16 token grid and a 768-dimensional semantic space. DINOv2-B uses its native 14×14 token grid with a 768-dimensional semantic space. Qwen-ViT supports native-resolution inputs, uses a 16×16 patch size, and has a 1152-dimensional semantic space in our setting. For all backbones, the reconstruction-aware patch embedding P_r is initialized from the pretrained patch embedding P_s, while all Transformer blocks F_ϕ are frozen. The reconstruction projection W_r maps the reconstruction-aware stream to 128 channels, yielding a 896 unified representation for PR-SigLIP2 and PR-DINOv2 and a 1280 representation for PR-Qwen-ViT.

Reconstruction training. For image reconstruction, we use a ViT-XL-scale decoder with 28 layers, hidden width 1152, FFN width 4096, and 16 attention heads. Following the RAE training recipe, the reconstruction objective combines pixel-level and perceptual losses, and the decoder is further refined with an adversarial loss for high-fidelity decoding. For SigLIP2-B and DINOv2-B, we first train P_r, W_r, and the decoder using L2 and LPIPS losses. We use AdamW with learning rate 2×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0, global batch size 512, one warmup epoch, and EMA decay 0.9978. We then freeze the encoder-side modules and train only the decoder for another 20 epochs with L2, LPIPS, and GAN losses, using the same optimizer, batch size, and EMA setting. The discriminator learning rate follows a cosine schedule from 2×10^{-4} to 2×10^{-5}. The main ImageNet results additionally use a short decoder calibration stage: we sample latents from an early DiT checkpoint and update only the decoder, while keeping the encoder and generative model fixed, so that the decoder adapts to realistic generation errors in the unified latent space. SigLIP2-B and DINOv2-B are trained on ImageNet-1K at 256×256 resolution, while Qwen-ViT is trained on web data with the same unified-tokenizer design.

Table 8: Comparison of visual interfaces in representative MoT-based unified models. "Spaces" denotes the number of distinct visual representation spaces used by the multimodal model.

Model	Visual interface	Pretrained semantic prior	Separate VAE	Spaces	Reference condition	Interface training
BAGEL (Deng et al., 2025)	SigLIP2 + FLUX-VAE	✓	✓	2	ViT + clean VAE tokens	Frozen
SenseNova-U1 (SenseNova-U1 Team, 2026)	Native pixels	–	–	1	Clean pixel tokens	End-to-end
UniSpace (Ours)	PR-Qwen-ViT tokenizer	✓	–	1	Unified T_u tokens	Frozen

Generation Training For ImageNet generation, we evaluate PR-SigLIP2 and PR-DINOv2 following the RAE (Zheng et al., 2025) generation protocol, adopting the same DiTwDDTHead architecture with input size 16, patch size 1, 896 input channels, hidden sizes $(1152, 2048)$, depths $(28, 2)$, and 16 attention heads. We use linear velocity prediction with a logit-normal time distribution and the dimension-dependent time-shift rule from RAE (Zheng et al., 2025); for the $896 \times 16 \times 16$ latent, the shift uses dimension 229,376 with base 4096. The model is trained with AdamW using learning rate 2×10^{-4}, betas $(0.9, 0.95)$, weight decay 0, global batch size 1024 with gradient accumulation 2, EMA decay 0.9995, gradient clipping 1.0, and a linear schedule with 40 warmup epochs decaying to 2×10^{-5} by epoch 800. Sampling uses an Euler ODE solver with 50 steps.

B Extended UniSpace Architecture and Training Details

Having established a representation that jointly preserves semantic and pixel-level information, we next investigate whether it can serve as the common visual interface of a large-scale multimodal model. We instantiate this idea as **UniSpace**, a unified multimodal model that performs text-to-image generation, instruction-based image editing, and image understanding in one patch-reparameterized visual representation space. Editing is especially important in this setting because it stress-tests both sides of the representation: the model must understand the reference image and instruction, perform the requested modification, and preserve fine-grained details in all irrelevant regions. UniSpace therefore uses the same frozen unified visual tokenizer for all three tasks and scales it within a Mixture-of-Transformer-Experts architecture.

B.1 Unified MoT Architecture

UniSpace builds on a decoder-only Qwen3-8B (Yang et al., 2025) backbone and adopts the Mixture-of-Transformer-Experts (MoT) design of BAGEL (Deng et al., 2025). It contains an understanding expert and a generation expert that operate on a common multimodal sequence. Token types are hard-routed to modality-specific parameters: text and conditioning-image tokens are processed by the understanding expert, whereas noised visual tokens to be predicted are processed by the generation expert. At every layer, self-attention allows tokens handled by the two experts to interact directly, retaining a bottleneck-free context across understanding and generation.

The key distinction of UniSpace lies in its visual interface. Existing MoT-based models typically use two visual representation spaces: a semantic ViT supplies tokens to the understanding pathway, while a separate VAE supplies latents to the generation pathway. In contrast, UniSpace uses the unified encoder and decoder developed in Sec. 3 as its only visual tokenizer, requires neither a separate VAE encoder nor an additional VAE latent space. Table 8 positions this design relative to two representative MoT systems. BAGEL retains two pretrained visual spaces for semantic conditioning and generation, while SenseNova-U1 removes external visual encoders and learns a native pixel interface end-to-end. UniSpace takes a different route: it preserves the semantic prior of a pretrained ViT, reparameterizes it into a reconstruction-compatible visual tokenizer, and consolidates understanding, generation, and editing into one frozen, reusable visual representation without introducing a separate VAE pathway.

For image understanding, an input image is encoded as T_u and processed together with the text instruction to predict the textual response. For text-to-image generation, the target image is represented in the same T_u space, corrupted along the flow path, and predicted by the generation expert conditioned on text. Image editing combines the two cases: the reference image is encoded as T_u^{ref} and provided together with the editing instruction, while the noised target representation is predicted in the same T_u space. Through shared self-attention, the target tokens condition on both the

instruction and the reference representation, and the sampled representation is decoded by the same decoder D_u. Thus, the three tasks differ in token arrangement and routing, but share the same visual tokenizer, representation space, and image decoder, as shown in Fig. 4.

B.2 Unified Training Objectives

UniSpace uses two training objectives over a shared multimodal sequence: cross-entropy next-token prediction for textual responses and balanced flow matching for visual prediction in the unified representation space T_u. For understanding samples, the model predicts response tokens with the standard next-token prediction objective

$$\mathcal{L}_{\mathrm{NTP}} = -\sum_{k=1}^{L} \log p_\theta(y_k \mid y_{<k}, c_{\mathrm{und}}), \tag{17}$$

where $y_{1:L}$ denotes the target textual response and c_{und} denotes the multimodal context constructed from the input image representation T_u and the text instruction. For generation and editing tasks, the target image is first mapped into the unified representation space as Z_1, and then perturbed following the linear flow path in Eq. 12. We sample the initial noise from a standard Gaussian distribution and define the target velocity as the displacement between the noise and target representation. Following RAE (Zheng et al., 2025), we adopt a logit-normal timestep sampling strategy with the same dimension-dependent time-shift rule. The model is optimized with the balanced component-wise flow matching objective in Eq. 15, where $\lambda_r = 0.75$ is used across all UniSpace training stages following the tokenizer-level generation setting.

The resulting unified visual prediction objective is

$$\mathcal{L}_{\mathrm{vis}}(c) = \mathbb{E}\left[\ell_{\mathrm{BFM}}\left(v_\theta(Z_t, t, c), V_t\right)\right]. \tag{18}$$

Here, c denotes the task-specific conditioning. For text-to-image generation, the condition contains only text tokens. For image editing, it additionally incorporates the reference image representation and the editing instruction. Thus,

$$\mathcal{L}_{\mathrm{t2i}} = \mathcal{L}_{\mathrm{vis}}(c_{\mathrm{t2i}}), \qquad \mathcal{L}_{\mathrm{edit}} = \mathcal{L}_{\mathrm{vis}}(c_{\mathrm{edit}}). \tag{19}$$

Generation and editing therefore share the same target representation space and visual prediction loss, differing only in their conditioning context.

Overall, UniSpace is trained on a mixture of understanding, text-to-image generation, and image-editing samples. Let $\tau \in \{\mathrm{und}, \mathrm{t2i}, \mathrm{edit}\}$ denote the task type of a training sample. The total training objective can be written as

$$\begin{aligned}\mathcal{L}_{\mathrm{UniSpace}} = \mathbb{E}_{(x,\tau)\sim\mathcal{D}}\big[&\mathbf{1}_{\tau=\mathrm{und}}\mathcal{L}_{\mathrm{NTP}} + \mathbf{1}_{\tau=\mathrm{t2i}}\mathcal{L}_{\mathrm{vis}}(c_{\mathrm{t2i}}) \\ &+ \mathbf{1}_{\tau=\mathrm{edit}}\mathcal{L}_{\mathrm{vis}}(c_{\mathrm{edit}})\big].\end{aligned} \tag{20}$$

Thus, understanding is supervised by cross-entropy over textual responses, while generation and editing share the same balanced flow-matching objective in the unified representation space T_u, differing only in their conditioning context. The unified encoder and decoder remain frozen throughout UniSpace training; all MoT parameters are trainable.

visual tokenizer remains frozen throughout.

B.3 Training Details

Training Data and Curriculum. We train UniSpace on internally curated datasets for text-to-image generation, image editing, and visual understanding. These data provide supervision for image generation, instruction-guided editing, and image-text understanding, respectively. The complete training curriculum is summarized in Table 9. We progressively scale training across three stages with increasing image resolutions of 256, 512, and 1024.

In Stage 1, we jointly optimize text-to-image generation and visual understanding with a sampling ratio of 10:1. In Stage 2, we introduce image editing and mix generation, editing, and understanding

Table 9: Progressive training curriculum of UniSpace. T2I, Edit, Und., and VLM denote text-to-image generation, image editing, visual understanding, and visual instruction tuning, respectively. Sample instances refer to the number of examples processed during training rather than the number of unique examples.

Item	Stage 1	Stage 2	Stage 3	SFT
Resolution	256	512	1024	1024
Training tasks	T2I + Und.	T2I + Edit + Und.	T2I + Edit + Und.	T2I + Edit + VLM
Sampling ratio	10:1	10:3:1	10:3:1	10:3:2
Optimization steps	170K	115K	60K	12K
Sample instances	264.2M	189.9M	50.9M	4.9M
Training tokens	110B	196B	151B	15B
Peak learning rate	2×10^{-4}	1×10^{-4}	4×10^{-5}	2×10^{-5}
Maximum sequence length	3,072	8,192	12,288	13,000
Hardware	256 NPUs	256 NPUs	256 NPUs	128 NPUs

data with a sampling ratio of 10:3:1. Stage 3 further scales all three capabilities through high-resolution packed multimodal training at 1024 resolution. Following pretraining, we perform a multimodal supervised fine-tuning (SFT) stage using packed text-to-image, image-editing, and LLaVA-NeXT-style visual instruction data (Liu et al., 2024b). During SFT, generation, editing, and visual instruction samples are mixed with a sampling ratio of 10:3:2. This stage improves instruction following and visual understanding while preserving the generation and editing capabilities acquired during pretraining.

Across all training stages, UniSpace processes approximately 510M training sample instances, corresponding to approximately 470B multimodal tokens. Here, sample instances refer to examples observed during training and may include repeated draws from the underlying datasets. The unified visual tokenizer and visual encoder remain frozen throughout all stages.

Optimization. We optimize UniSpace using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-15}$, and zero weight decay. We use a linear warmup followed by a constant learning-rate schedule. The warmup length is 600 steps for the three pretraining stages and 100 steps for SFT. The peak learning rates for each stage are reported in Table 9. We train in BF16 precision and clip the global gradient norm to 1.0. We use one packed sequence per device at each optimization step without gradient accumulation. No exponential moving average is used during training.

The model is trained progressively: each stage is initialized from the selected checkpoint of the preceding stage, while the optimizer and learning-rate scheduler are reinitialized at the beginning of each new stage. The visual encoder and the unified visual tokenizer are frozen throughout training, while the language model, modality experts, and multimodal projection modules are optimized jointly.

Packed Multimodal Training. To improve training efficiency under heterogeneous multimodal sequence lengths, we employ packed multimodal training. Multiple independent examples, potentially from different training objectives, are concatenated into a single packed sequence until a predefined token budget is reached. Samples are drawn according to the task-specific sampling ratios in Table 9. Consequently, the specified ratios describe the expected sampling frequencies over the full training run rather than the exact composition of every individual packed sequence.

We construct block-wise attention masks for packed sequences to prevent information exchange between different examples. In particular, tokens may only attend to tokens belonging to the same example, while attention within each example follows its corresponding modality-specific pattern. Text tokens use causal attention, whereas visual tokens use full attention within the same sample. Position indices are reset for every individual example inside a packed sequence. Losses are computed only over valid target tokens: the text cross-entropy loss is applied to target text tokens, while the image generation loss is applied only to target visual tokens.

The maximum packed sequence lengths are $3{,}072$, $8{,}192$, $12{,}288$, and $13{,}000$ for Stage 1, Stage 2, Stage 3, and SFT, respectively. The resulting token utilization is approximately 73% to 79% across stages. As the image resolution increases, each packed sequence contains fewer examples due to the increased number of visual tokens; however, the global task sampling distribution remains unchanged.

Data Processing. We use resolution-specific image buckets for the three pretraining stages. Images are resized and bucketed according to their aspect ratios, with an aspect-ratio deviation threshold of 0.05. Visual inputs are tokenized using the frozen unified visual tokenizer. For image generation, we apply text conditioning dropout with probability 0.1 to support classifier-free guidance. We also apply instruction dropout with probability 0.1 during multimodal training. For image editing, reference images are processed through the understanding branch, while the target image is represented as a generation target.

Distributed Training Infrastructure. We train Stages 1–3 on 256 Ascend 910B NPUs with 64GB memory each, distributed over 16 nodes with 16 NPUs per node. The SFT stage is trained on 128 Ascend 910B NPUs over 8 nodes. Training is implemented in PyTorch using FSDP with the `SHARD_GRAD_OP` sharding strategy. We use BF16 mixed precision, activation checkpointing, and SDPA-based attention to reduce memory consumption. The selected training trajectory requires approximately 142K NPU-hours in total.

Table 10: Reconstruction quality on the ImageNet-1K 256×256 validation set. "Ratio" denotes the spatial downsampling ratio.

Method	Type	Training Data	Ratio	PSNR↑	SSIM↑	rFID↓
Tokenizers without Demonstrated Semantic Capability						
LlamaGen (Sun et al., 2024)	Discrete-Pixel	MS+IN-1K	16	20.65	0.54	2.47
Open-MAGVIT2 (Luo et al., 2024)	Discrete-Pixel	Mixed-100M	16	22.70	0.64	1.67
SD-VAE XL (Rombach et al., 2021)	Continuous-Pixel	OImg+LAae++	8	27.37	0.78	0.67
Qwen-Image (Wu et al., 2025a)	Continuous-Pixel	–	8	**32.18**	0.90	1.45
SD-VAE 3 (Rombach et al., 2021)	Continuous-Pixel	–	8	31.29	0.87	0.20
Wan2.1 (Wan et al., 2025)	Continuous-Pixel	–	8	31.34	0.89	0.95
FLUX-VAE (Labs, 2024)	Continuous-Pixel	–	8	32.74	**0.92**	**0.18**
VA-VAE (Yao et al., 2025b)	Continuous-Pixel	IN-1K	16	27.96	0.79	0.28
Wan2.2 (Wan et al., 2025)	Continuous-Pixel	–	16	31.25	0.88	0.74
Tokenizers with Demonstrated Semantic Capability						
Show-o (Xie et al., 2025)	Discrete-Pixel	–	16	21.34	0.59	3.50
QLIP-B (Zhao et al., 2025)	Discrete-Pixel	DC-1B	16	23.16	0.63	3.21
VILA-U (Wu et al., 2025c)	Discrete-Pixel	WL-10B+CY-1B	16	–	–	1.80
TokenFlow (Qu et al., 2025)	Discrete-Pixel	LA+CY	16	21.41	0.69	1.37
UniTok (Ma et al., 2026)	Discrete-Pixel	DC-1B	16	27.28	0.77	0.41
UniLIP (Tang et al., 2025)	Continuous-Pixel	BP-32M	32	22.99	0.75	0.79
BLIP3-o (Chen et al., 2025a)	Continuous-Diffusion	BP-32M	16	14.71	0.58	3.18
VTP-L (Yao et al., 2025a)	Continuous-Pixel	DC-277M	16	25.82	0.74	0.36
UniFlow (SigLIP2) (Yue et al., 2025)	Continuous-Diffusion	IN-1K	16	29.38	0.93	0.62
UniFlow (DINOv2) (Yue et al., 2025)	Continuous-Diffusion	IN-1K	14	**31.01**	**0.94**	0.54
RAE (SigLIP2-B) (Zheng et al., 2025)	Continuous-Pixel	IN-1K	16	19.35	0.49	0.53
RAE (DINOv2-B) (Zheng et al., 2025)	Continuous-Pixel	IN-1K	14	18.86	0.48	0.57
RAEv2 (DINOv3-L, K=7)† (Singh et al., 2026)	Continuous-Pixel	IN-1K	16	22.57	0.63	0.29
PR-SigLIP2	Continuous-Pixel	IN-1K	16	29.64	0.87	0.18
PR-DINOv2	Continuous-Pixel	IN-1K	14	30.84	0.90	**0.14**
PR-Qwen-ViT	Continuous-Pixel	Web Data	16	30.16	0.88	0.17

Note. † For RAEv2, we report the K=7 variant, which is used as its reconstruction–generation trade-off setting. VTP-L reports semantic capability through ImageNet zero-shot and linear probing evaluation.

C Extended Experimental Results

C.1 Unified Tokenizer Evaluation

For reconstruction evaluation, we use the ImageNet-1K validation set to assess PR-SigLIP2 and PR-DINOv2, reporting PSNR, SSIM, and rFID as evaluation metrics. For multimodal understanding, we follow the LLaVA-v1.5 (Liu et al., 2024a) setup, using Vicuna-7B-v1.5 as the language model and a two-layer MLP as the multimodal projector. The visual encoder is frozen, and the complete unified representation T_u is provided as input to the projector. For ImageNet generation, we report gFID, sFID, IS, precision, and recall in Table 12.

C.1.1 High-Fidelity Image Reconstruction

We first assess whether Patch Reparameterization equips pretrained semantic encoders with high-fidelity image reconstruction capability. As shown in Table 10, the proposed method consistently achieves strong reconstruction quality across PR-SigLIP2, PR-DINOv2, and PR-Qwen-ViT. In particular, PR-DINOv2 attains an rFID of 0.14, a PSNR of 30.84, and an SSIM of 0.90, achieving the best rFID among the compared tokenizers with demonstrated semantic capability. Figure 8 presents qualitative comparisons on representative ImageNet validation images, with red boxes highlighting fine local details and text-like regions that are particularly sensitive to reconstruction quality.

The matched-backbone comparisons further demonstrate the effectiveness of Patch Reparameterization. Compared with RAE using the corresponding SigLIP2-B and DINOv2-B encoders, PR-SigLIP2 reduces rFID from 0.53 to 0.18, while PR-DINOv2 reduces it from 0.57 to 0.14, corresponding to relative reductions of 66.0% and 75.4%, respectively. For PR-DINOv2, PSNR increases from 18.86 to 30.84 and SSIM from 0.48 to 0.90. Our method also compares favorably with the

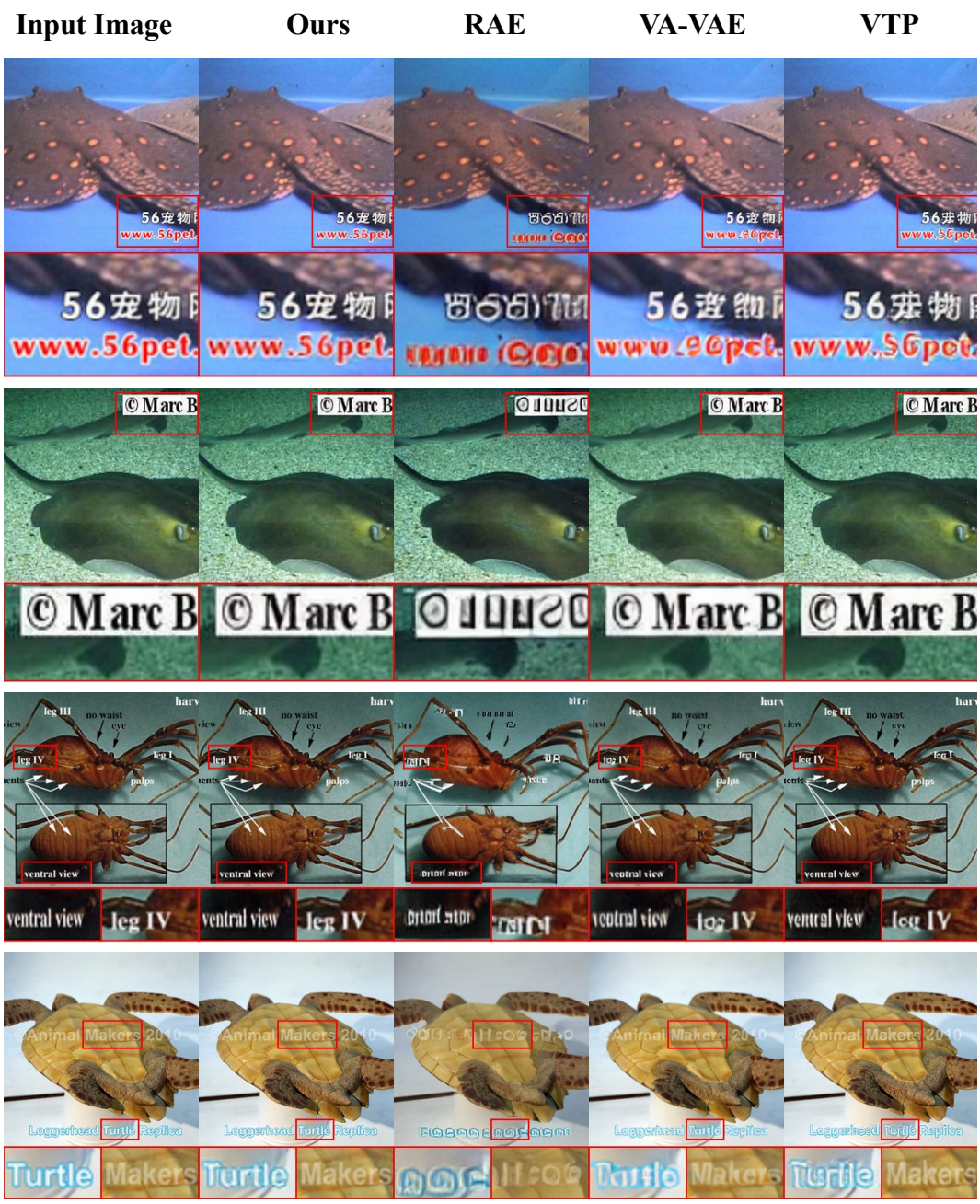

Figure 8: Qualitative reconstruction comparison. Columns show the input image, Ours, RAE, VA-VAE, and VTP. Red boxes highlight local details and text-like regions.

Table 11: Multimodal understanding performance of unified visual tokenizers. Our patch-reparameterized tokenizers are evaluated using the complete unified representation T_u. Methods marked with † use the LLaVA-v1.5 (Liu et al., 2024a) training data. The "Original Semantic Encoders" rows are matched baselines using the same semantic encoder, LLM, training data, and evaluation protocol as our corresponding variants.

Method	Visual Encoder	LLM	Res.	POPE	GQA	TQA	MMV	MMB	MME-S	MME-P	Avg.
Existing Unified Visual Tokenizers											
VILA-U † (Wu et al., 2025c)	SigLIP-SO400M	Vicuna-7B	256	81.6	–	–	–	–	–	1311.6	–
UniTok † (Ma et al., 2026)	ViTamin-L	Vicuna-7B	256	81.7	–	–	–	–	–	1448.0	–
QLIP † (Zhao et al., 2025)	CLIP-L	Vicuna-7B	392	86.1	61.8	55.2	33.3	–	–	1498.3	–
TokenFlow-B † (Qu et al., 2025)	CLIP-B	Vicuna-13B	224	84.0	59.3	49.8	22.4	55.3	1660.4	1353.6	60.21
TokenFlow-L † (Qu et al., 2025)	ViTamin-XL	Vicuna-13B	256	85.0	60.3	54.1	27.7	60.3	1622.9	1365.4	62.40
UniTok (Ma et al., 2026)	ViTamin-L	LLaMA-2-7B	256	83.2	61.1	51.6	33.9	–	–	1448.0	–
TokLIP (Lin et al., 2025b)	VQ-GAN+ViT-SO400M	Qwen2.5-7B	384	84.1	59.5	–	29.8	67.6	–	1448.4	–
TokenFlow-XL (Qu et al., 2025)	SigLIP-SO400M	Qwen2.5-14B	384	87.8	62.5	62.3	48.2	76.8	1922.2	1551.1	73.04
UniFlow-LV † (Yue et al., 2025)	DFN-CLIP-L	Vicuna-7B	224	86.56	61.38	53.40	30.2	63.83	1748.0	1446.9	65.02
UniFlow-LV † (Yue et al., 2025)	SigLIP2-SO400M	Vicuna-7B	256	87.94	63.29	58.0	32.4	68.38	1823.0	1477.9	67.87
UniFlow-LV † (Yue et al., 2025)	DINOv2-L	Vicuna-7B	378	88.04	59.37	45.53	25.6	51.48	1590.5	1257.7	58.92
UniFlow-LV † (Yue et al., 2025)	InternViT-300M (Chen et al., 2024)	Vicuna-7B	448	88.97	63.35	61.85	36.6	67.10	1803.0	1505.1	69.04
Original Semantic Encoders											
SigLIP2 baseline †	SigLIP2-B	Vicuna-7B	256	85.2	61.4	54.36	25.0	64.4	1689.3	1378.2	63.39
Qwen-ViT baseline †	Qwen-ViT	Vicuna-7B	448	86.26	63.25	64.27	31.88	69.24	1791.3	1471.3	68.29
Ours: Patch-Reparameterized Unified Visual Tokenizers											
PR-SigLIP2 †	PR-SigLIP2	Vicuna-7B	256	84.8	61.1	54.6	25.5	65.2	1741.1	1447.2	64.37
PR-Qwen-ViT †	PR-Qwen-ViT	Vicuna-7B	448	86.57	63.13	63.68	30.28	69.07	1874.5	1522.7	68.94

Note: The average is computed as $\text{Avg} = (\text{POPE} + \text{GQA} + \text{TQA} + \text{MMV} + \text{MMB} + \text{MME-S}/20 + \text{MME-P}/20)/7$.

recent RAEv2 reconstruction-enhanced variant: PR-DINOv2 improves PSNR from 22.57 to 30.84 and rFID from 0.29 to 0.14, while using the smaller DINOv2-B encoder rather than the DINOv3-L backbone used by RAEv2. These improvements are achieved with the pretrained ViT backbone kept frozen, indicating that its Transformer blocks can effectively propagate reconstruction-relevant visual details when driven by a reconstruction-aware patch embedding.

The proposed representations are also competitive with tokenizers specifically designed for pixel reconstruction and generation. PR-DINOv2 improves over VA-VAE by reducing rFID from 0.28 to 0.14, and even surpasses large-scale generative VAEs such as FLUX-VAE and SD-VAE 3 in terms of rFID. Meanwhile, PR-Qwen-ViT shows that Patch Reparameterization generalizes beyond ImageNet-pretrained encoders. Crucially, unlike reconstruction-specialized tokenizers, our representations retain the semantic capabilities of their respective pretrained encoders, as evaluated in the next section.

C.1.2 Multimodal Understanding

We next examine whether Patch Reparameterization compromises the semantic understanding inherited from the pretrained encoder. We compare against the corresponding original semantic encoders under the same LLaVA-v1.5 setup, using identical language models, multimodal projectors, training data, and evaluation protocols. In all experiments, the complete unified representation T_u, rather than the semantic component T_s alone, is fed to the multimodal projector.

As shown in Table 11, Patch Reparameterization does not compromise the overall multimodal understanding capability of the original encoders. PR-SigLIP2 achieves an average score of 64.37, surpassing the original SigLIP2-B baseline of 63.39, while PR-Qwen-ViT reaches 68.94, slightly above the original Qwen-ViT baseline of 68.29. Beyond the matched-backbone comparisons, our method also achieves highly competitive understanding performance under the LLaVA-v1.5/Vicuna-7B setting. These results demonstrate that Patch Reparameterization retains the semantic capability of pretrained encoders while incorporating reconstruction-aware visual details into the unified representation.

Table 12: System-level reconstruction–generation performance on ImageNet 256×256 with representative latent generative priors, following the evaluation protocols reported by VA-VAE (Yao et al., 2025b) and RAE (Zheng et al., 2025).

		Reconstruction	Model	Generation w/o CFG					Generation w/ CFG				
Method	Tokenizer	rFID↓	#Params	gFID↓	sFID↓	IS↑	Pre.↑	Rec.↑	gFID↓	sFID↓	IS↑	Pre.↑	Rec.↑
Representative Latent Generative Priors													
REPA (Yu et al., 2024)	SD-VAE	0.61	675M	5.90	–	–	–	–	1.42	4.70	305.7	0.80	0.65
LightningDiT (Yao et al., 2025b)	VA-VAE	0.28	675M	2.17	4.36	205.6	0.77	0.65	1.35	4.15	295.3	0.79	0.65
UniFlow (MAR) (Yue et al., 2025)	UniFlow (InternViT)	0.28	479M	2.45	–	228.0	–	–	1.85	–	290.0	–	–
VTP-L (LightningDiT) (Yao et al., 2025a)	VTP-L	0.36	675M	1.85	–	232.3	0.79	0.63	1.11	–	279.5	0.79	0.67
RAE (DiT^{DH}-XL) (Zheng et al., 2025)	RAE (DINOv2-B)	0.57	839M	1.51	–	242.9	0.79	0.63	1.13	–	262.6	0.78	0.67
RAEv2 (DiT^{DH}-XL) (Singh et al., 2026)	RAEv2 (DINOv3-L, K=7)	0.29	0.9B	1.65	–	228.0	–	–	1.06	–	255.3	–	–
Ours: Patch-Reparameterized Unified Visual Tokenizers													
PR-SigLIP2	PR-SigLIP2	0.18	839M	4.42	6.66	190.4	0.72	0.65	2.80	6.05	248.2	0.78	0.61
PR-DINOv2	PR-DINOv2	0.14	839M	2.10	5.39	217.2	0.78	0.64	1.87	4.89	274.3	0.82	0.60

Note. For RAEv2, we report the K=7 reconstruction–generation trade-off setting from the original paper. "–" indicates metrics not reported in the corresponding source under the matched setting.

C.1.3 ImageNet Generation

More importantly, our method establishes a favorable reconstruction–generation operating point for latent generative modeling. As shown in Table 12, PR-DINOv2 shifts the trade-off toward substantially higher reconstruction fidelity while retaining effective generation performance, achieving an rFID of 0.14 and gFIDs of 2.10 without classifier-free guidance and 1.87 with classifier-free guidance at scale 1.2. In contrast, RAE, RAEv2, and VTP-L achieve stronger generation results but have substantially higher tokenizer rFID values of 0.57, 0.29, and 0.36, respectively. Thus, Patch Reparameterization recovers substantially more decoder-relevant visual detail from a pretrained semantic ViT while still producing a latent space that can be effectively modeled by a generative prior.

Taken together with the reconstruction and multimodal understanding results, these findings demonstrate that the proposed representation is not merely a reconstruction tokenizer or a generation latent. Instead, it provides a unified visual representation that retains multimodal understanding, supports high-fidelity reconstruction, and remains effective for latent image generation.

C.1.4 Auxiliary Diagnostic Study on Entangled Unified Representations

As an auxiliary diagnostic experiment, separate from our final design, we investigate whether jointly supporting semantic understanding and high-fidelity reconstruction is alone sufficient for generative modeling. We construct a deliberately entangled representation by merging semantic and reconstruction features into a single latent space without explicit decomposition. As shown in Fig. 13, this diagnostic variant performs well when evaluated on real encoded latents, achieving a zero-shot accuracy of 78.53 (compared with 79.10 for the SigLIP baseline), an rFID of 0.069, and a PSNR of 33.83.

However, the entangled representation fails to support faithful generative decoding. After training a DiT prior on the merged representation, decoding the generated latent $\widehat{Z}_m$ with the pretrained high-fidelity reconstruction decoder D_r results in an FID of 120.9, indicating a severe generation failure. Interestingly, decoding the same $\widehat{Z}_m$ with the semantic decoder D_s yields a much lower FID of 8.07. This result does not indicate successful generation in the unified latent space; rather, it shows that the generative prior mainly captures the dominant semantic structure, while failing to preserve the reconstruction-relevant information required by D_r.

We further find that approximately 95% of the merger output variation is explained by the semantic pathway. This semantic dominance explains the decoder-dependent generation gap: although the generated latent retains enough semantic structure for D_s to produce plausible images, it does not preserve the fine-grained directions needed for high-fidelity decoding by D_r. These results reveal a gap between representation quality and generative modelability: good reconstruction on real encoded latents does not guarantee faithful reconstruction after generation. This motivates our explicit factorization, in which T_s and $\widetilde{T}_r$ are concatenated along the channel dimension, making the reconstruction component directly accessible during flow-matching training. Further details of the entangled representation and diagnostic protocol are provided in Appendix F.

Table 13: Quantitative comparison on ImgEdit. All category scores and the overall score are higher-is-better. Baseline values are reproduced from the evaluation reported by SenseNova-U1 (SenseNova-U1 Team, 2026). "Active" denotes activated generation parameters for sparse models.

Model	# Params	Add	Adjust	Extract	Replace	Remove	Bg.	Style	Hybrid	Action	Overall↑
Closed-source Models											
UniWorld-V2 (Li et al., 2025)	–	4.29	4.44	**4.32**	**4.69**	**4.72**	4.41	4.91	3.83	4.83	**4.49**
Nano-Banana-Pro	–	4.44	4.62	3.42	4.60	4.63	4.32	**4.97**	3.64	4.69	4.37
Seedream 4.5	–	4.57	**4.65**	2.97	4.66	4.46	4.37	4.92	3.71	4.56	4.32
Seedream 4.0	–	4.33	4.38	3.89	4.65	4.57	4.35	4.22	3.71	4.61	4.30
Nano-Banana	–	**4.62**	4.41	3.68	4.34	4.39	4.40	4.18	3.72	4.83	4.29
GPT-Image-1	–	4.61	4.33	2.90	4.35	3.66	**4.57**	4.93	**3.96**	**4.89**	4.20
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
Open-source Image Generation and Editing Models											
Qwen-Image-Edit-2511 (Wu et al., 2025a)	20B	**4.54**	**4.57**	4.13	4.70	4.46	4.36	4.89	**4.16**	4.81	**4.51**
LongCat-Image-Edit (Team et al., 2025)	6B	4.44	4.53	3.83	**4.80**	4.60	4.33	**4.92**	3.75	**4.82**	4.45
FLUX.2 [Dev] (Labs, 2024)	32B	4.50	4.18	3.83	4.65	**4.65**	4.31	4.88	3.46	4.70	4.35
Qwen-Image-Edit-2509 (Wu et al., 2025a)	20B	4.32	4.36	4.04	4.64	4.52	4.37	4.84	3.39	4.71	4.35
Z-Image-Edit (Cai et al., 2025)	6B	4.40	4.14	**4.30**	4.57	4.13	4.14	4.85	3.63	4.50	4.30
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
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
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
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
OmniGen (Xiao et al., 2025)	3.8B	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
Open-source Unified Multimodal Models											
Emu3.5 (Cui et al., 2025)	32B	**4.61**	4.32	**3.96**	**4.84**	**4.58**	**4.35**	4.79	**3.69**	4.57	**4.41**
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
SenseNova-U1 (SenseNova-U1 Team, 2026)	8B	3.83	4.15	3.12	4.32	3.26	4.18	4.85	3.03	4.41	3.90
InternVL-U (w/ CoT) (Tian et al., 2026)	1.7B	4.24	3.80	2.58	4.36	3.51	3.92	4.69	3.00	4.31	3.82
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
UniWorld-V1 (Lin et al., 2025a)	12B	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
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
UniSpace (Ours)	8B	4.53	**4.38**	3.61	4.67	4.42	4.23	4.55	2.70	4.47	4.28

Table 14: Additional image-editing evaluation on GEdit. Reference GEdit-Bench-EN results are reproduced from SenseNova-U1 (SenseNova-U1 Team, 2026); BAGEL-CN is reproduced from the BAGEL (Deng et al., 2025) report; UniSpace is evaluated with GPT-4o. Avg. is computed over English and Chinese overall scores when both are available. All metrics are higher-is-better.

		GEdit-Bench-EN			GEdit-Bench-CN			
Model	# Params	SC	PQ	Overall↑	SC	PQ	Overall↑	Avg.↑
Specialized Image Editing Models								
Qwen-Image-Edit-2511 (Wu et al., 2025a)	20B	**8.30**	**8.20**	**7.88**	–	–	–	–
LongCat-Image-Edit (Team et al., 2025)	6B	8.13	8.18	7.75	–	–	–	–
Z-Image-Edit (Cai et al., 2025)	6B	8.11	7.72	7.57	–	–	–	–
Qwen-Image-Edit (Wu et al., 2025a)	20B	8.00	7.86	7.56	–	–	–	–
Unified Multimodal Models								
Emu3.5 (Cui et al., 2025)	32B	8.11	**7.70**	**7.59**	–	–	–	–
SenseNova-U1 (SenseNova-U1 Team, 2026)	8B	8.27	7.49	7.47	–	–	–	–
BAGEL (Deng et al., 2025)	7B	7.36	6.83	6.52	7.34	6.85	6.50	6.51
UniSpace (Ours)	8B	**8.29**	7.06	7.41	**8.27**	**7.00**	**7.38**	**7.395**

C.2 Scaling Unified Multimodal Modeling

We next study whether the proposed unified representation can be scaled from controlled representation learning to a full multimodal system. UniSpace is not intended to isolate the representation alone: it combines the frozen unified encoder–decoder with an 8B Mixture-of-Transformer-Experts

backbone and is trained on large-scale internal generation, editing, and understanding data. We therefore use UniSpace as a system-level validation. The central question is whether a single ViT-based visual space can replace the conventional separation between semantic image tokens and VAE latents while still supporting practical generation and editing.

C.2.1 Image Editing

We evaluate UniSpace on two image-editing benchmarks, ImgEdit (Ye et al., 2025) and GEdit (SenseNova-U1 Team, 2026). ImgEdit serves as our primary benchmark for system-level comparison, while GEdit provides a complementary evaluation of instruction following and perceptual quality under both English and Chinese prompts.

ImgEdit. ImgEdit covers nine editing categories, ranging from local content manipulation to style transfer and hybrid editing. As shown in Table 13, UniSpace achieves an overall score of 4.28 at the 8B scale. It substantially outperforms comparable-scale unified models, including SenseNova-U1 (3.90) and BAGEL (3.20), while approaching the 4.41 score of the substantially larger 32B Emu3.5. These results demonstrate the strong editing capability of UniSpace at the 8B scale. Despite using substantially fewer parameters than Emu3.5, UniSpace achieves a comparable overall score, while clearly outperforming other unified models of similar scale. This suggests that the unified representation provides sufficient semantic and reconstruction information for precise instruction-based image editing.

GEdit. We further evaluate UniSpace on GEdit. As GEdit serves as a complementary benchmark, we report aggregate results and compare UniSpace with other representative methods in Table 14. UniSpace achieves overall scores of 7.41 and 7.38 on the English and Chinese full sets, respectively, yielding a bilingual average of 7.395. It outperforms BAGEL and remains close to the dense SenseNova-U1 8B model. The detailed results show that UniSpace achieves strong semantic consistency, while its perceptual quality remains relatively weaker, particularly on compositional editing and other perceptually sensitive cases.

Qualitative Results. Figure 9 presents qualitative comparisons on four representative editing tasks: object addition, style transfer, scene replacement, and object substitution. UniSpace consistently follows the editing instructions while preserving the identity and structure of unrelated image regions. In contrast, BAGEL and SenseNova-U1 tend to produce artifacts or exhibit incomplete instruction adherence in challenging cases, particularly for fine-grained style transfer and multi-object compositional editing. Figure 10 shows additional results on human-centric editing tasks, including expression change, accessory addition, clothing color modification, and hairstyle transfer. These examples require fine-grained localized editing while preserving identity and unrelated appearance attributes. UniSpace performs these edits while maintaining the overall appearance and structural consistency of the input subjects, further supporting its ability to combine instruction following with localized visual editing.

C.2.2 Text-to-Image Generation

We evaluate UniSpace on three text-to-image generation benchmarks: GenEval (Ghosh et al., 2023), OneIG-Bench (Chang et al., 2025), and DPG-Bench (Hu et al., 2024). GenEval evaluates object-centric compositional alignment, OneIG-Bench measures fine-grained generation across English and Chinese prompts, and DPG-Bench evaluates dense-prompt following across global, entity, attribute, and relational constraints. We adopt the comparison sets collected by SenseNova-U1 (SenseNova-U1 Team, 2026) and report the results in Tables 15–18.

GenEval. As shown in Table 15, UniSpace achieves an overall score of 0.84. UniSpace performs well on single-object generation, two-object generation, positional alignment, and attribute binding, although its counting and color-binding scores remain relatively weaker.

OneIG-Bench. On OneIG-Bench, UniSpace obtains overall scores of 0.561 and 0.533 on the English and Chinese subsets, respectively, yielding a bilingual average of 0.547. As shown in Tables 16 and 17, UniSpace achieves the strongest bilingual average among the compared unified multimodal

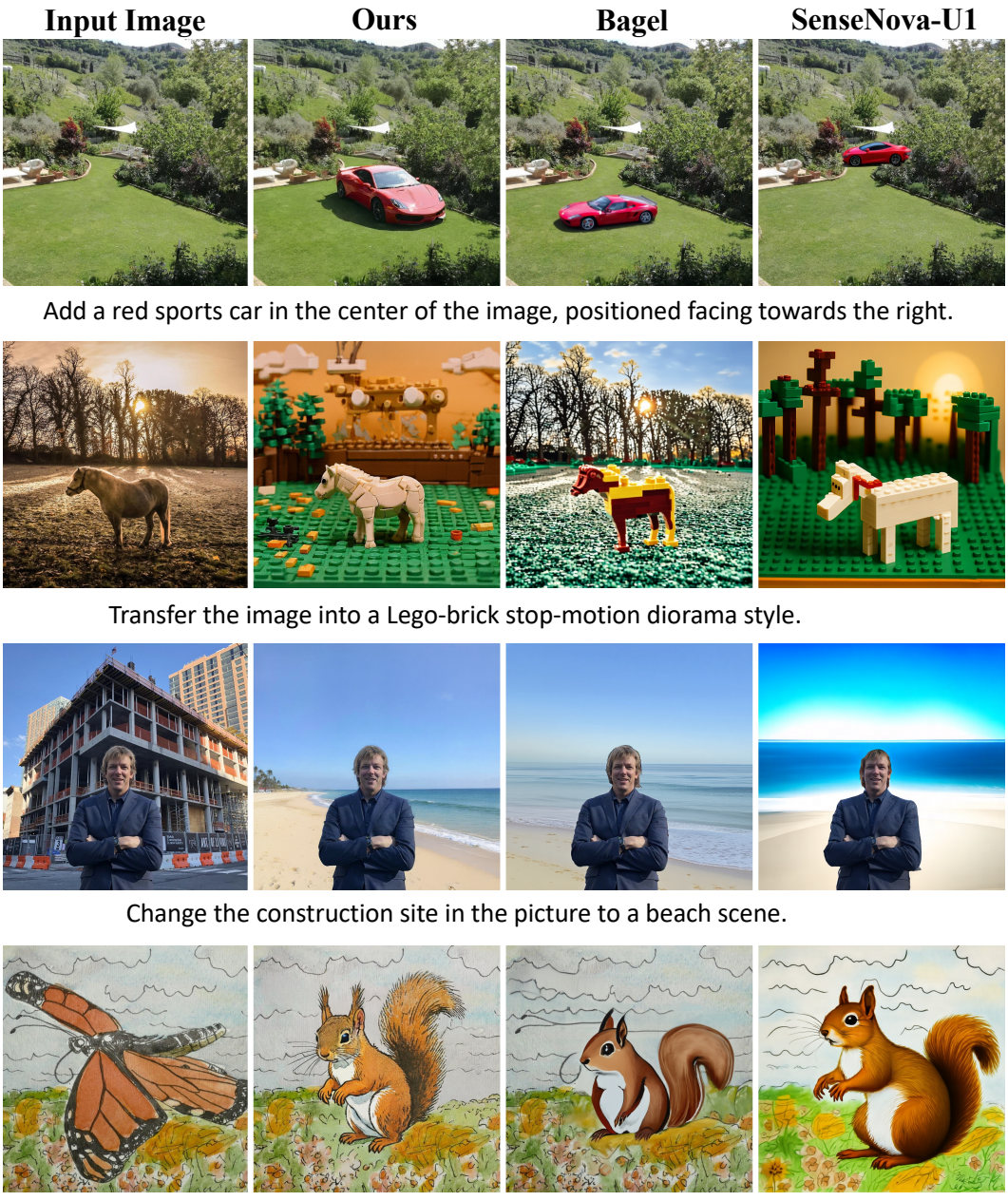

Figure 9: Qualitative image editing comparison. Columns (left to right): input image, Ours (UniSpace), BAGEL, SenseNova-U1. Rows (top to bottom): object addition, style transfer, background replacement, object replacement.



Figure 10: Qualitative comparison on human-centric image editing. Columns (left to right): input image, Ours (UniSpace), BAGEL, SenseNova-U1. Rows: expression change (neutral→smile), accessory addition (graduation cap), clothing recoloring (blue→green robes), hairstyle transfer (braids→curly), expression change (smile→serious).

Table 15: Text-to-image generation on GenEval. SO, TO, CT, CL, POS, and ATTR denote single object, two objects, counting, colors, position, and attribute binding, respectively. All metrics are higher-is-better.

Model	# Params	SO	TO	CT	CL	POS	ATTR	Overall↑
Open-source Generation Models								
Qwen-Image (Wu et al., 2025a)	20B	0.99	0.92	**0.89**	0.88	**0.76**	**0.77**	**0.87**
Z-Image (Cai et al., 2025)	6B	**1.00**	**0.94**	0.78	**0.93**	0.62	**0.77**	0.84
SD3-Medium (Rombach et al., 2021)	2B	0.99	**0.94**	0.72	0.89	0.33	0.60	0.74
FLUX.1-dev (Labs, 2024)	12B	0.98	0.81	0.74	0.79	0.22	0.45	0.66
Open-source Unified Multimodal Models								
SenseNova-U1 (SenseNova-U1 Team, 2026)	8B	**1.00**	0.96	**0.92**	**0.92**	**0.91**	0.76	**0.91**
InternVL-U (Tian et al., 2026)	1.7B	0.99	0.94	0.74	0.91	0.77	0.74	0.85
BAGEL (Deng et al., 2025)	7B	0.99	0.94	0.81	0.88	0.64	0.63	0.82
Janus-Pro (Chen et al., 2025b)	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80
OmniGen2 (Wu et al., 2025b)	4B	**1.00**	0.95	0.64	0.88	0.55	0.76	0.80
UniWorld-V1 (Lin et al., 2025a)	12B	0.99	0.93	0.79	0.89	0.49	0.70	0.80
Show-o2 (Xie et al., 2026)	7B	**1.00**	0.87	0.58	**0.92**	0.52	0.62	0.76
Emu3.5 (Cui et al., 2025)	32B	–	–	–	–	–	–	0.73
UniSpace (Ours)	8B	0.98	0.92	0.69	0.88	0.83	0.73	0.84

models, slightly exceeding Emu3.5 (0.546) and SenseNova-U1 (0.542). It also obtains the highest style scores among the compared methods on both the English and Chinese subsets, with scores of 0.467 and 0.455, respectively. These results demonstrate that UniSpace remains competitive in multilingual text-to-image generation, particularly in style control and fine-grained visual synthesis.

DPG-Bench. On DPG-Bench, UniSpace achieves an overall score of 86.49. It outperforms BAGEL (85.07), Show-o2 (86.14), and most other listed unified multimodal models. UniSpace achieves the highest relation score among all compared models, reaching 94.97, and obtains a strong entity score of 92.26. These results indicate that UniSpace can effectively follow dense prompts, particularly those involving relational and entity-level constraints, although its global and attribute scores remain below those of the strongest text-to-image systems.

Qualitative Results. Figures 11 and 12 present qualitative text-to-image comparisons between UniSpace, BAGEL, and SenseNova-U1. UniSpace generally produces images that are well aligned with the input prompts while maintaining coherent object layouts, attributes, and visual styles. Compared with the other unified models, its generations exhibit a more realistic appearance and richer visual details, especially in complex scenes and fine-grained visual content. These results provide qualitative evidence that the unified representation supports both prompt alignment and high-fidelity visual synthesis.

C.2.3 Unified Capability Evaluation

The same model also retains an understanding pathway, but we do not use UniSpace's system-level understanding scores as the main evidence for this paper. UniSpace is optimized primarily for generation and editing, and its post-training recipe is not designed to maximize standard VLM benchmarks. Instead, the understanding capability of the unified representation itself is validated in the controlled experiments in Table 11, where all tasks use the complete unified representation T_u.



Figure 11: Qualitative text-to-image comparison. Columns show Ours (UniSpace), BAGEL, and SenseNova-U1. Each row is generated from the prompt shown below the corresponding images.

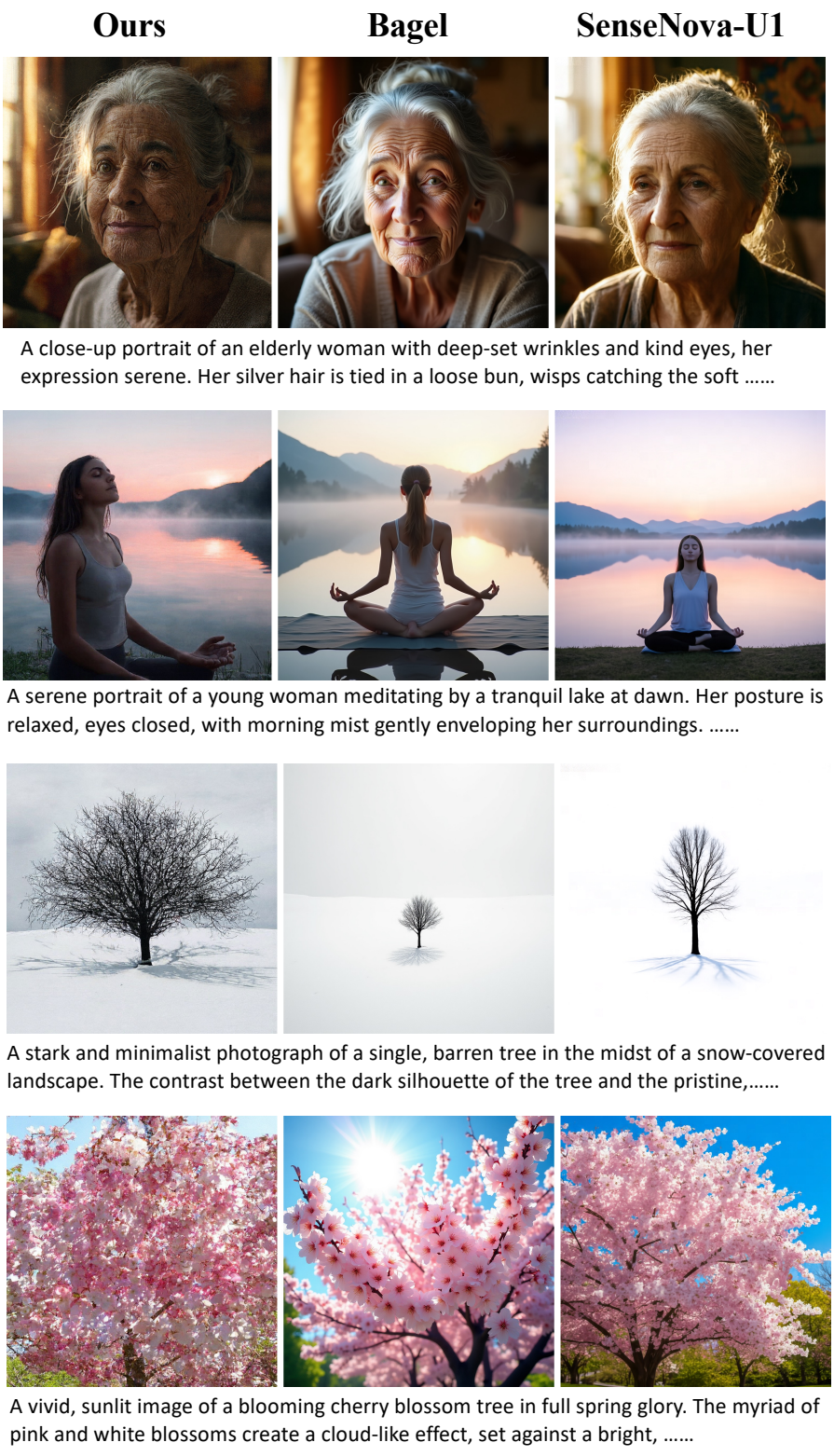

Figure 12: Additional qualitative text-to-image comparison. Columns show Ours (UniSpace), BAGEL, and SenseNova-U1. Each row is generated from the prompt shown below the corresponding images.

Table 16: Text-to-image generation on OneIG-Bench (English).

Model	# Params	Align.	Text	Reason.	Style	Div.	Overall↑
Closed-source Models							
Gemini-2.5-Flash-Image	–	**0.878**	**0.894**	**0.346**	0.450	0.182	**0.550**
GPT-Image-1	–	0.851	0.857	0.345	**0.462**	0.151	0.533
Seedream 3.0	–	0.818	0.865	0.275	0.413	**0.277**	0.530
Open-source Generation Models							
Qwen-Image (Wu et al., 2025a)	20B	0.882	0.891	0.306	0.418	0.197	0.539
Open-source Unified Multimodal Models							
Emu3.5 (Cui et al., 2025)	32B	**0.902**	**0.994**	**0.345**	0.427	0.151	**0.564**
SenseNova-U1 (SenseNova-U1 Team, 2026)	8B	0.882	0.969	0.330	0.396	0.166	0.549
BAGEL (Deng et al., 2025)	7B	0.769	0.244	0.173	0.367	0.251	0.361
Janus-Pro (Chen et al., 2025b)	7B	0.553	0.001	0.139	0.276	**0.365**	0.267
UniSpace (Ours)	8B	0.860	0.937	0.311	**0.467**	0.233	0.561

Table 17: Text-to-image generation on OneIG-Bench (Chinese). Baseline results are reproduced from SenseNova-U1 (SenseNova-U1 Team, 2026). All metrics are higher-is-better.

Model	# Params	Align.	Text	Reason.	Style	Div.	Overall↑
Closed-source Models							
Gemini-2.5-Flash-Image	–	**0.825**	0.276	0.298	0.427	0.198	0.337
GPT-Image-1	–	0.812	0.650	**0.300**	**0.449**	0.159	0.474
Seedream 3.0	–	0.793	**0.928**	0.281	0.397	**0.243**	**0.528**
Open-source Generation Models							
Qwen-Image (Wu et al., 2025a)	20B	0.825	0.963	0.267	0.405	0.279	0.548
Open-source Unified Multimodal Models							
Emu3.5 (Cui et al., 2025)	32B	**0.853**	0.941	0.300	0.386	0.166	0.529
SenseNova-U1 (SenseNova-U1 Team, 2026)	8B	0.826	**0.977**	**0.303**	0.392	0.176	**0.535**
BAGEL (Deng et al., 2025)	7B	0.672	0.365	0.186	0.357	0.268	0.370
Janus-Pro (Chen et al., 2025b)	7B	0.324	0.148	0.104	0.264	**0.358**	0.240
UniSpace (Ours)	8B	0.807	0.881	0.276	**0.455**	0.244	0.533

Table 18: Dense-prompt following performance on DPG-Bench. All metrics are higher-is-better.

Model	# Params	Global	Entity	Attribute	Relation	Other	Overall↑
Closed-source Models							
Seedream 4.5	–	89.24	**94.30**	**92.14**	92.23	**93.83**	**88.63**
Nano-Banana-Pro	–	**91.00**	92.85	91.56	92.39	89.93	87.16
GPT-Image-1	–	88.89	88.94	89.84	**92.63**	90.96	85.15
Open-source Generation Models							
Qwen-Image (Wu et al., 2025a)	20B	91.32	**91.56**	92.02	**94.31**	**92.73**	**88.32**
Z-Image (Cai et al., 2025)	6B	**93.39**	91.22	**93.16**	92.22	91.52	88.14
SD3-Medium (Rombach et al., 2021)	2B	87.90	91.01	88.83	80.70	88.68	84.08
FLUX.1-dev (Labs, 2024)	12B	74.35	90.00	88.96	90.87	88.33	83.84
Open-source Unified Multimodal Models							
SenseNova-U1 (SenseNova-U1 Team, 2026)	8B	88.74	90.90	**92.43**	92.43	**92.50**	**87.78**
Tuna (Liu et al., 2026)	7B	90.42	91.68	90.94	91.87	90.73	86.76
NEO-unify (SenseNova-U1 Team, 2026)	8B	**91.00**	91.53	92.06	94.14	90.43	86.71
Show-o2 (Xie et al., 2026)	7B	–	–	–	–	–	86.14
InternVL-U (Tian et al., 2026)	1.7B	90.39	90.78	90.68	90.29	88.77	85.18
BAGEL (Deng et al., 2025)	7B	88.94	90.37	91.29	90.82	88.67	85.07
Janus-Pro (Chen et al., 2025b)	7B	86.90	88.90	89.40	89.32	89.48	84.19
Ovis-U1 (Wang et al., 2025)	1.2B	82.37	90.08	88.68	93.35	85.20	83.72
OmniGen2 (Wu et al., 2025b)	4B	88.81	88.83	90.18	89.37	90.27	83.57
UniWorld-V1 (Lin et al., 2025a)	12B	83.64	88.39	88.44	89.27	87.22	81.38
UniSpace (Ours)	8B	84.80	**92.26**	90.00	**94.97**	88.80	86.49

D Ablation on Balanced Flow Matching

We ablate the reconstruction-component objective weight λ_r under the same $768{+}128$ unified representation used by our final model. All variants share the same encoder, decoder, training schedule, and sampling setting, isolating the effect of the semantic–reconstruction balance in the flow-matching objective.

The sweep shows a clear middle optimum rather than a monotonic preference for larger reconstruction weights. Increasing the reconstruction objective share from $\lambda_r = 0.25$ to $\lambda_r = 0.75$ improves FID from 8.99 to 7.10, indicating that under-weighting the reconstruction component leaves decoder-critical visual details insufficiently modeled. However, further increasing the reconstruction share does not continue to improve generation, with $\lambda_r = 0.92$ producing 7.39. This supports assigning sufficient but not overwhelming objective weight to the reconstruction component. The best FID is obtained at $\lambda_r = 0.75$, matching the weighting used in our final balanced flow-matching objective.

Table 19: Ablation of reconstruction-component weighting in balanced flow matching. All variants use the same $768{+}128$ unified representation and are evaluated under the same controlled generation setting.

Recon. share λ_r	FID↓	IS↑	Precision↑
0.25	8.99	142.4	0.681
0.60	7.42	153.2	0.696
0.75	**7.10**	**154.4**	0.704
0.82	7.24	152.8	**0.705**
0.86	7.13	152.5	0.704
0.92	7.39	150.0	0.702

E Ablation on Reconstruction Compression

We next study the channel compression in the Channel Factorized Merge. This ablation compares our default 128-dimensional reconstruction component with a no-compression variant that preserves the full 768 reconstruction channels. Both variants use the same semantic component and the same factorized concatenation design; only the dimensionality of $\widetilde{T}_r$ is changed.

The no-compression variant confirms the expected reconstruction benefit of a larger reconstruction stream: increasing d_r from 128 to 768 improves rFID from 0.163 to 0.085 and PSNR from 29.79 to 33.61. However, this extra detail comes at a clear cost for generation. At both 20 and 40 epochs, the full-dimensional reconstruction component is substantially harder for the DiT prior to model, with FID degrading from 9.95 to 15.51 at epoch 20 and from 6.92 to 11.07 at epoch 40. These results support the role of compression in the Channel Factorized Merge: it is not intended to maximize reconstruction in isolation, but to retain sufficient decoder-critical details while reducing the complexity of the reconstruction distribution that the generative model must learn.

Table 20: Ablation of reconstruction-component compression. Reducing the reconstruction stream from 768 to 128 channels slightly weakens pure reconstruction metrics, but substantially improves generative modeling under the same early training budget.

Recon. dim d_r	Unified dim	Epoch	PSNR↑	SSIM↑	rFID↓	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
128	768+128	20	29.79	0.873	0.163	9.95	7.73	117.5	0.713	0.565
128	768+128	40	29.79	0.873	0.163	**6.92**	**6.93**	**144.4**	**0.730**	0.587
768	768+768	20	**33.61**	**0.940**	**0.085**	15.51	8.45	90.0	0.655	0.574
768	768+768	40	**33.61**	**0.940**	**0.085**	11.07	7.80	115.2	0.679	**0.596**

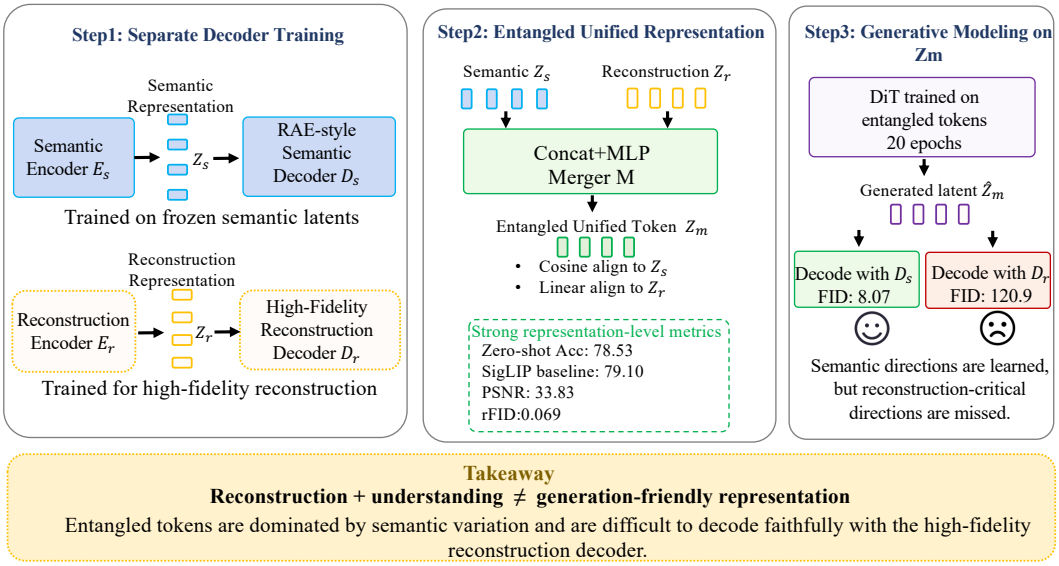

Figure 13: Diagnostic study of an entangled unified representation. We construct an MLP-merged token $Z_m = M(Z_s, Z_r)$ from semantic and reconstruction representations and align it with both. Although Z_m supports strong understanding and high-fidelity reconstruction on real encoded latents, its generative decoding fails when using the pretrained high-fidelity reconstruction decoder D_r, which yields an FID of 120.9. Decoding the same generated latent with the semantic decoder D_s gives a much lower FID of 8.07, indicating that the generative prior captures semantic structure but fails to preserve reconstruction-relevant information in the entangled latent space.

F Entangled Representation and Diagnostic Protocol

We provide additional details of the diagnostic experiment in Sec. C.1.4. Let I denote an input image, and let

$$Z_s = E_s(I), \qquad Z_r = E_r(I)$$

denote the outputs of a pretrained semantic encoder and a reconstruction-oriented encoder, respectively. We merge the two representations using an MLP:

$$Z_m = M([Z_s; Z_r]),$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation. Unlike the factorized representation used in our method, Z_m does not preserve an explicit semantic–reconstruction decomposition.

The merger is trained with the same semantic and reconstruction alignment objectives described in the main text. Specifically, Z_m is aligned with Z_s to preserve semantic information, while a linear projection of Z_m is aligned with Z_r to retain reconstruction-relevant information. The semantic decoder D_s and the high-fidelity reconstruction decoder D_r are trained separately before the diagnostic experiment, with their corresponding encoders kept frozen. At the representation level, the resulting Z_m simultaneously supports semantic understanding and high-fidelity reconstruction on real encoded latents. It achieves a zero-shot accuracy of 78.53, compared with 79.10 for the SigLIP baseline, together with a PSNR of 33.83 and an rFID of 0.069.

We next train a DiT prior directly on Z_m for 20 epochs using the same image distribution and evaluation protocol as in the main experiments. The resulting generated latent exhibits a pronounced decoder-dependent gap: decoding it with the high-fidelity reconstruction decoder D_r yields an FID of 120.9, whereas decoding it with the semantic decoder D_s yields a much lower FID of 8.07. This gap indicates that the DiT prior primarily learns the semantic-dominant variation in the entangled representation, while failing to model the reconstruction-relevant variation required by D_r. Thus, the fact that Z_m supports both semantic understanding and high-fidelity reconstruction on real encoded latents does not ensure that both types of information are equally accessible to a generative prior. The reconstruction information is present, but its entangled composition with semantic variation makes it difficult for the generative model to identify and control.

To quantify the relative contribution of the semantic pathway, we measure the fraction of the merger output variation explained by the semantic input:

$$\rho_s = \frac{\mathrm{Var}(M(Z_s, 0))}{\mathrm{Var}(M(Z_s, Z_r))}.$$

The variance is computed over the evaluation image set using the same merger parameters. We obtain $\rho_s \approx 95\%$, indicating that most of the merger output variation is explained by the semantic pathway. This suggests that reconstruction-relevant directions, although sufficient for reconstructing real encoded latents, occupy a relatively small and difficult-to-control part of the entangled latent space. This observation motivates the explicit factorization used in our final representation, where T_s and $\widetilde{T}_r$ are concatenated along the channel dimension and remain directly accessible during flow-matching training.