UNICOM: Unified Multimodal Modeling via Compressed Continuous Semantic Representations

Yaqi Zhao[1,2,4]* **Wang Lin**[3,4,*] **Zijian Zhang**[4] **Miles Yang**[4]
Zhao Zhong[4] **Liefeng Bo**[4] **Jingyuan Chen**[3†] **Wentao Zhang**[1,†]

[1]Peking University
[2]The Hong Kong University of Science and Technology
[3]Zhejiang University [4]Tencent Hunyuan

Website **Code** **Models**

Abstract

Current unified multimodal models typically rely on discrete visual tokenizers to bridge the modality gap. However, discretization inevitably discards fine-grained semantic information, leading to suboptimal performance in visual understanding tasks. Conversely, directly modeling continuous semantic representations (e.g., CLIP, SigLIP) poses significant challenges in high-dimensional generative modeling, resulting in slow convergence and training instability. To resolve this dilemma, we introduce **UNICOM**, a unified framework that harmonizes multimodal understanding and generation via compressed continuous semantic representations. We empirically demonstrate that channel-wise compression is significantly more effective than spatial token reduction, retaining competitive VAE-free reconstruction fidelity while substantially improving generative convergence. Accordingly, we design an attention-based semantic compressor to distill dense features into a compact unified representation. A Transfusion-style predictor then generates these compressed latents with dense spatial correspondence. At scale, UNICOM achieves competitive text-to-image generation and state-of-the-art performance on complex image editing without relying on VAE latents. The gains are especially clear on knowledge-intensive benchmarks.

1 Introduction

The pursuit of unified multimodal models [44, 13, 57], capable of both understanding and generating content across modalities such as text and images, represents a significant paradigm shift in artificial intelligence. A key insight from this evolution is that unification is not merely an engineering challenge of connecting modules. It fundamentally questions the representational divide between discrete, symbolic reasoning (language) and continuous, perceptual synthesis (vision). The ongoing research is, in essence, searching for a "unified token", a common representational substrate that can seamlessly flow between understanding and generation, between text and pixels, while being efficiently processed by a scalable neural architecture.

Achieving such unification critically depends on how visual information is represented, making the design of visual encoders a central challenge. Some compromise approaches [10, 60, 47, 71, 13] adopt hybrid encoders that combine variational autoencoder (VAE [49]) latents with Visual Transformer (ViT) features. While this design facilitates generation, it introduces an inherent representational divergence: semantic understanding and image synthesis are grounded in different feature

*Equal contribution. Work done during internships at Tencent Hunyuan.
†Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

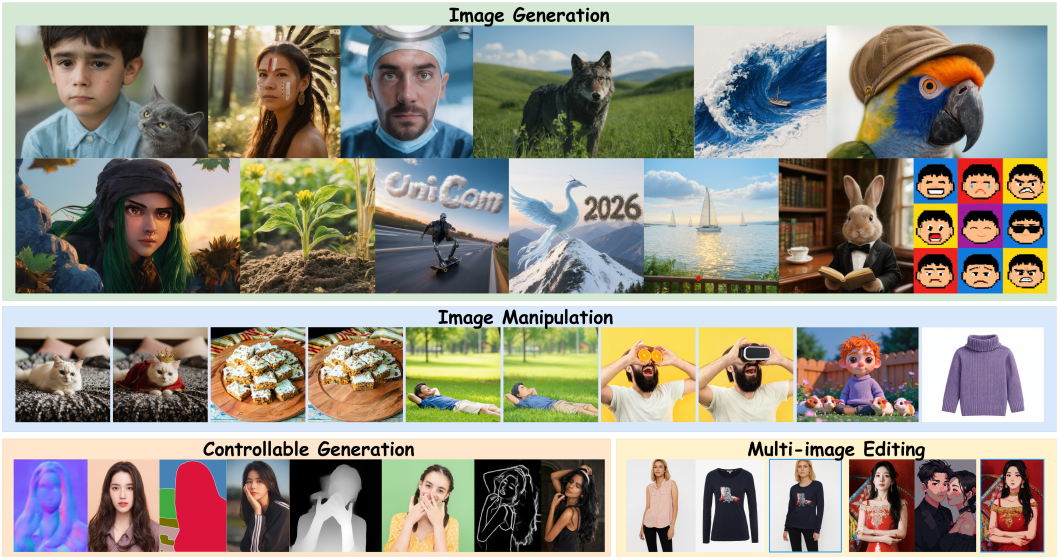


Figure 1: **High-quality samples generated by UNICOM**. Built on compressed continuous representations, our unified model demonstrates exceptional capabilities in text-to-image generation, precise image manipulation, multi-image editing, and fine-grained controllable generation.

spaces, fundamentally limiting deeper unification. To address this representational schism, some works [20, 39, 23] move toward using only ViT-based visual embeddings, aiming to anchor both understanding and generation within a single semantic space. Within this line of research, some approaches discretize powerful continuous features (*e.g.*, CLIP/SigLIP [48, 58]) via vector quantization to obtain tractable visual tokens. Although discretization simplifies generative modeling, it inevitably incurs irreversible information loss, particularly in fine-grained spatial and textural details that are essential for high-fidelity image synthesis.

More recent methods [82, 57, 19, 9], abandon quantization altogether and instead operate directly on continuous ViT representations. However, this shift exposes a new challenge: the high-dimensional feature manifold is redundant, curved, often non-smooth, and difficult to model generatively (Fig 5). As illustrated in Fig 2 (left), we therefore learn a compressor $\mathcal{C}_\phi : \mathcal{Z} \to \tilde{\mathcal{Z}}$ that maps the original semantic space $\mathcal{Z} \subset \mathbb{R}^D$ to a compact chart $\tilde{\mathcal{Z}} \subset \mathbb{R}^d$. This compression should dynamically balance low-frequency semantic structure for understanding and high-frequency visual details for generation, while removing redundant directions. We thus revisit visual embeddings from two perspectives: **modeling** and **prediction**.

For visual embedding modeling, prior unified systems commonly reduce semantic features by spatial downsampling [6, 52, 34], which shortens the target sequence but merges image regions and weakens local correspondence. We instead find that compressing along the channel dimension better preserves semantic fidelity and pixel-level details. We therefore use the channel-compressed configuration ($n = 1024, d = 64$) as our generative target. This is implemented by an attention-based semantic compressor that maps high-dimensional visual semantics into a compact continuous latent space, which preserves essential information while simplifying the target distribution for generative modeling. Empirically, this representation preserves visual understanding (Tab 5), achieves the strongest reconstruction metrics among comparable VAE-free semantic representation methods (Tab 3), and substantially improves training stability and efficiency (Fig 9).

For visual embedding prediction, we adopt a Transfusion [83] architecture to predict compressed visual latents over mixed-modality sequences. This full-sequence formulation preserves dense correspondence between the condition and target image, which is critical for faithful editing. Compared with predicting uncompressed features, the compact target enables more stable and efficient training, reaching comparable generation quality with roughly 3,458 fewer GPU hours and 18.1M fewer training samples (Fig 5). We also analyze a query-based alternative [45, 6] and find that its query bottleneck converges more slowly and weakens structural consistency.

Finally, we propose **UNICOM**, a unified multimodal model that performs generation directly over compressed visual embeddings. UNICOM achieves competitive text-to-image generation and is particularly strong on editing tasks that require semantic reasoning and visual consistency. Specifically, it leads open-source models on KRIS-Bench and WorldEdit, and achieves the highest Consistency Score on GEdit-Bench, even without relying on VAE latents. These results show that semantically rich and well-compressed visual embeddings can serve as an effective universal interface for both understanding and generation. Our contributions are as follows:

- We establish an effective paradigm for unifying visual understanding and generation by learning to predict continuous, compressed semantic embeddings, which we demonstrate preserves both high-level semantics and fine-grained visual details better than prior quantization methods.
- We show that channel compression is significantly superior to sequence reduction for preserving information, and that an attention-based compressor is crucial for maintaining semantic structure.
- The proposed UNICOM model validates this paradigm at scale, achieving state-of-the-art or competitive performance across image reconstruction, text-to-image generation, and challenging image editing tasks.

2 Related Work

A prominent approach attempts [54, 61, 44, 46, 79] to integrate image generation into the autoregressive framework of LLMs, addressing the challenge of image tokenization. Early methods [20, 22, 39, 11], serialized vision-language encoders like CLIP/SigLIP into token sequences similar to text, typically paired with a diffusion decoder to combine AR understanding with diffusion-based generation. However, the resulting quantization loss and long sequences often lacked the semantic alignment necessary for efficient cross-modal reasoning. Some approaches [1, 53, 6], like MetaQuery [45], generate query tokens via an LLM, which are then reconstructed to compact visual representations for image decoding through a diffusion model. Another line of work [50, 71, 40, 57], such as Transfusion [83], combines AR's sequential reasoning with diffusion's high-fidelity image synthesis. In these models, text tokens are generated autoregressively by an LLM, while image tokens undergo denoising conditioned on the textual context.

3 UNICOM

3.1 Problem Formulation and Unified Framework

Our work posits that the key to a unified and efficient model lies not in circumventing the VLM's semantic space, but in *transforming it* into a form amenable to generative modeling while preserving its rich informational content. We formulate a two-stage generative process that decomposes the conditional image distribution $P(\mathbf{x} \mid \mathbf{c})$ (where $\mathbf{c}$ is a conditioning context such as text) as follows:

$$P(\mathbf{x} \mid \mathbf{c}) = \int P(\tilde{\mathbf{z}} \mid \mathbf{c}) \cdot P(\mathbf{x} \mid \tilde{\mathbf{z}})\, d\tilde{\mathbf{z}}. \tag{1}$$

Here, $\mathbf{Z} \in \mathcal{Z}$ denotes the original high-dimensional semantic feature of an image, while $\tilde{\mathbf{z}} \in \tilde{\mathcal{Z}} \subset \mathbb{R}^{N \times d}$ (with $d \ll D$) denotes its compressed continuous latent. This formulation introduces a critical intermediate variable $\tilde{\mathbf{z}}$ with two properties:

1. **Semantic Fidelity:** $\tilde{\mathbf{z}}$ retains the necessary information from $\mathbf{Z}$ to reconstruct the image $\mathbf{x}$ with high fidelity, ensuring a strong connection to understanding.
2. **Generative Tractability:** The distribution $P(\tilde{\mathbf{z}} \mid \mathbf{c})$ over the lower-dimensional space $\tilde{\mathcal{Z}}$ is significantly smoother for learning and sampling.

Our framework, illustrated in Fig 2, operationalizes this formulation through three core components: Semantic Compressor §3.2, Generative Prior §3.3, and Diffusion Decoder. The compressor and decoder are first jointly pre-trained using a reconstruction objective to establish a well-behaved $\tilde{\mathcal{Z}}$ space. Subsequently, the compressor is frozen, and the Generative prior is trained to sample within this fixed, tractable space.

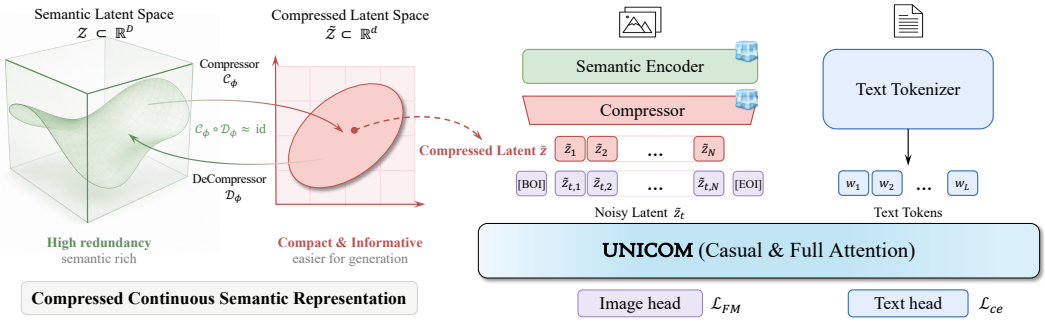


Figure 2: **Overview of UNICOM.** UNICOM compresses high-dimensional semantic features into a compact continuous latent space that preserves visual semantics while making generation tractable. This shared space serves as the unified interface where a Transfusion-style transformer predicts compressed visual latents and a diffusion decoder maps them back to images.

3.2 Semantic Compression and Reconstruction

The core of our framework is the construction of the compressed semantic latent space $\tilde{\mathcal{Z}}$. This module must reconcile two competing objectives: **minimizing information loss** to preserve the visual and semantic fidelity required for high-quality reconstruction, and **maximizing tractability** by simplifying the underlying data distribution for robust generative modeling. Formally, we seek a parametric function $\mathcal{C}_\phi : \mathcal{Z} \to \tilde{\mathcal{Z}}$ that performs a nonlinear dimensionality reduction, where $\tilde{\mathcal{Z}} \subset \mathbb{R}^{N \times d}$ and $d \ll D$.

Why Compression is Necessary. The native VLM embedding space $\mathcal{Z}$, while semantically rich, exhibits high dimensionality and a complex, often multi-modal and discontinuous, topology. Directly modeling the distribution $P(\mathbf{Z} \mid \mathbf{c})$ with a generative model (*e.g.*, a diffusion process) is empirically unstable and leads to high FID (Tab 4), as the model struggles to navigate the intricate geometry of $\mathcal{Z}$. This observation aligns with the manifold hypothesis; we posit that the information essential for generation lies on a lower-dimensional submanifold within $\mathcal{Z}$. The compressor $\mathcal{C}_\phi$ thus learns a compact parameterization of this submanifold.

Design of the Attention-Based Compressor. Prior work, such as VUGEN [9], employs a simple Multi-Layer Perceptron (MLP) for compression. While effective, an MLP processes each spatial feature token independently, potentially disregarding the crucial long-range contextual relationships between image patches that are inherently captured by the vision encoder's self-attention layers.

To preserve cross-token semantics during compression, we follow the attention projection design of UniTok [39]. Specifically, we replace the concatenation of attention-head outputs with averaging to reduce channels, and use concatenation for the reverse expansion:

$$\begin{aligned} \mathbf{H}_{\mathrm{c}}(\mathbf{Z}) &= \frac{1}{h}\sum_{i=1}^{h} \mathrm{Attention}_i^{\mathrm{c}}(\mathbf{Z}), \\ \mathbf{H}_{\mathrm{e}}(\tilde{\mathbf{z}}) &= \left[\mathrm{Attention}_1^{\mathrm{e}}(\tilde{\mathbf{z}}); \ldots; \mathrm{Attention}_h^{\mathrm{e}}(\tilde{\mathbf{z}})\right]. \end{aligned} \tag{2}$$

Here, each head produces $N \times d$ features, $h = D/d$, and brackets denote channel concatenation. Thus, $\mathbf{H}_{\mathrm{c}} \in \mathbb{R}^{N \times d}$ and $\mathbf{H}_{\mathrm{e}} \in \mathbb{R}^{N \times D}$ preserve the spatial token grid. Each attention branch is combined with a normalized linear projection branch and a residual feed-forward layer to form the compressor $\mathcal{C}_\phi$ and decompressor $\mathcal{E}_\eta$. Unlike UniTok, we retain continuous latents without vector quantization.

Joint Optimization with the Diffusion Decoder. The compressor is not trained in isolation. Its parameters ϕ are jointly optimized with the diffusion decoder parameters ψ using a reconstruction objective. Given an image $\mathbf{x}$, its visual features $\mathbf{Z} = f_{\text{und}}(\mathbf{x})$, and compressed latents $\tilde{\mathbf{z}} = \mathcal{C}_\phi(\mathbf{Z})$,

the decoder reconstructs the image as $\hat{\mathbf{x}} = \mathcal{D}_\psi(\tilde{\mathbf{z}})$. Training minimizes the composite loss:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{flow}}(\mathbf{x}, \hat{\mathbf{x}}) + \lambda \cdot \mathcal{L}_{\text{perc}}(\mathbf{x}, \hat{\mathbf{x}}), \tag{3}$$

where $\mathcal{L}_{\text{flow}}$ denotes the flow-matching loss used by the diffusion decoder and $\mathcal{L}_{\text{perc}}$ is a perceptual loss (e.g., LPIPS [81]).

This reconstruction objective encourages $\mathcal{C}_\phi$ to discard information redundant for image reconstruction while preserving semantic and appearance cues needed for generation. We therefore prioritize reconstruction fidelity under a fixed training budget, without imposing an additional contrastive or captioning loss on the compressor.

3.3 Unified Prediction with Transfusion

With $\tilde{\mathcal{Z}}$ established, UNICOM follows Transfusion [83] and models $P(\tilde{\mathbf{z}} \mid \mathbf{c})$ in a single, fully trainable transformer. The model operates on a unified sequence that interleaves discrete text tokens and continuous latent representations of images.

Given a text prompt $\mathbf{c}$, it is tokenized into a sequence of tokens $\{\mathbf{w}_1, \ldots, \mathbf{w}_L\}$. To generate an image, we prepend a special `[BOI]` (Beginning of Image) token and append a special `[EOI]` (End of Image) token to a placeholder sequence of N latent vectors initialized as pure noise $\tilde{\mathbf{z}}_T \sim \mathcal{N}(0, I)$. For visual inputs, the semantic encoder extracts $\mathbf{Z} = f_{\text{und}}(\mathbf{x})$: understanding uses $\mathbf{Z}$ directly to preserve fine-grained details, while editing uses $\tilde{\mathbf{z}} = \mathcal{C}_\phi(\mathbf{Z})$ to facilitate efficient multi-turn editing.

The mixed sequence uses causal attention for text and bidirectional attention among latent vectors within each image segment. This preserves dense spatial correspondence for editing. We analyze a query-guided alternative in Appendix E.

Flow Matching Training Objective. During training, given a text condition $\mathbf{c}$ and its corresponding ground-truth compressed semantic representation $\tilde{\mathbf{z}}_1$, we construct training pairs following the standard Flow Matching procedure. Specifically, we sample a time step $t \sim \mathcal{U}[0, 1]$ and noise $\epsilon \sim \mathcal{N}(0, I)$, and compute the interpolated latent

$$\tilde{\mathbf{z}}_t = t\tilde{\mathbf{z}}_1 + (1 - t)\epsilon, \tag{4}$$

along with the target velocity

$$\mathbf{v}_t = \tilde{\mathbf{z}}_1 - \epsilon. \tag{5}$$

The condition $\mathbf{c}$ and the noised latent $\tilde{\mathbf{z}}_t$ are assembled into the mixed input sequence and passed through the Transformer. The model predicts the velocity field from the output hidden states corresponding to the positions of $\tilde{\mathbf{z}}_t$. Optimization is performed using a mean squared error loss:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t, \mathbf{c}, \tilde{\mathbf{z}}_1, \epsilon} \left[\| \mathbf{v}_t - \mathbf{v}_\theta(\tilde{\mathbf{z}}_t, t; \mathbf{c}) \|_2^2 \right]. \tag{6}$$

During this stage, both the semantic compressor $\mathcal{C}_\phi$ and the diffusion decoder $\mathcal{D}_\psi$ are frozen. Consequently, the transformer learns to perform conditional sampling on a fixed, well-structured semantic manifold $\tilde{\mathcal{Z}}$.

4 Experiments

In this section, we evaluate UNICOM, a unified multimodal model built on compressed continuous semantic representations. We describe implementation details in Section 4.1, report main results in Section 4.2, and ablate the major components of unified multimodal modeling in Section 4.3, including representation shape, compressor architecture, and prediction pathway.

4.1 Implementation Details

Our unified model is built upon Qwen-2.5-7B-Instruct [55] for multimodal understanding and FLUX.1-dev [29] for high-fidelity image generation, utilizing SigLIP2-SO400M-Patch16-NaFlex [58] as the semantic encoder. To effectively bridge the modality gap, we employ a Multi-Head Attention (MHA) semantic compressor for feature compression and a two-layer MLP connector for visual-language alignment. The training pipeline consists of four progressive stages: alignment, pre-training, continued training, and supervised fine-tuning, all adopting a native-resolution

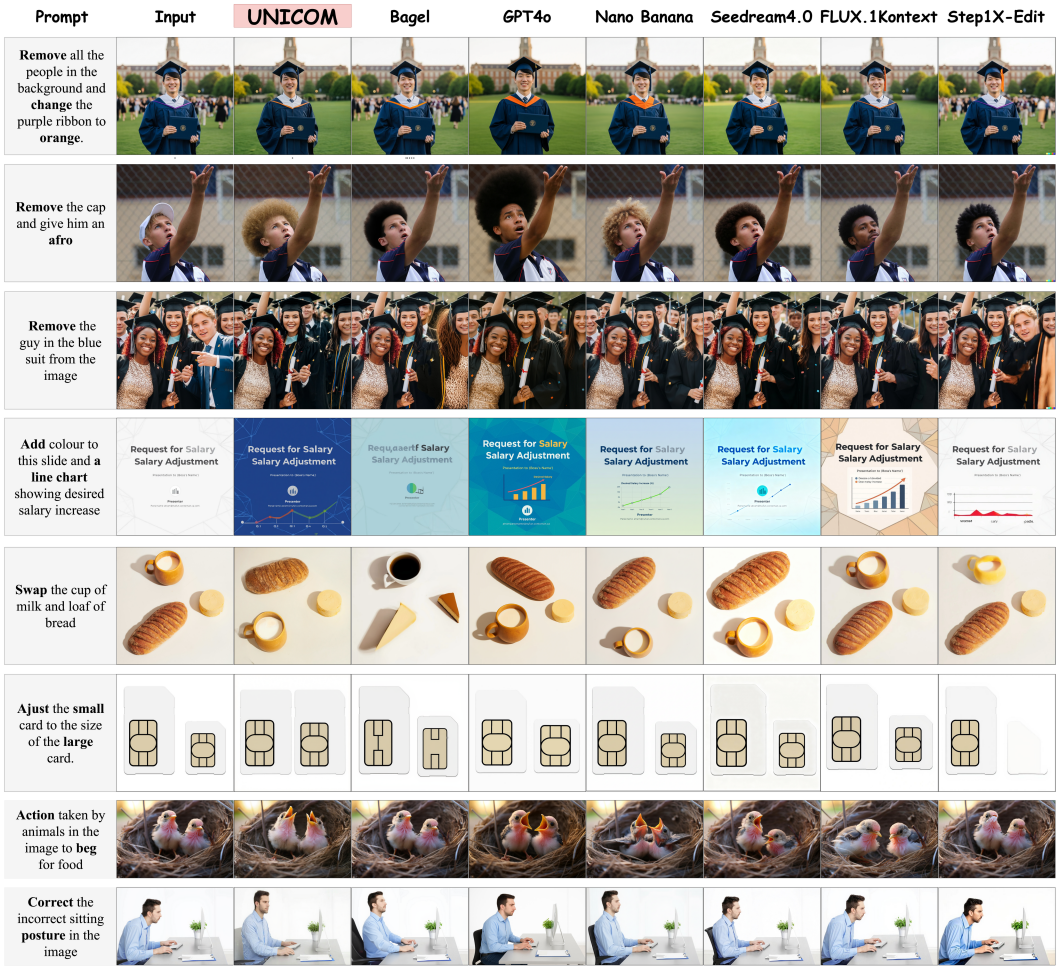


Figure 3: **Comparison of the results for image editing**, highlighting UNICOM performance in tasks such as replace, remove, add, swap and knowledge-intensive editing. See more in Appendix I.

strategy. For generative training and inference, we use a shifted noise schedule with shift $= 3.0$ by default unless otherwise specified. See Appendix A for the decoder configuration, hyperparameters, data mixtures, and ablation setup, and Appendix C for comparisons across backbones, encoders, and decoders.

4.2 Main Results

Text to Image Generation. As shown in Tab 1, on GenEval [21] and DPG-Bench [24], our method achieves competitive performance among unified models. UNICOM is particularly strong on the more challenging WISE [42] benchmark, where semantic priors are essential. We attribute this advantage primarily to directly using visual-semantic features such as SigLIP as the learning target. Compared with traditional VAE latent spaces, this semantically rich representation is more suitable for encoding and reconstructing textual semantic information.

Image Editing. As shown in Tab 2, our model also demonstrates strong performance in image editing tasks, achieving leading scores on ImgEdit-Bench [75] and GEdit-Bench [36]. Many previous works [57, 9, 19] that generate based on comprehension priors either rarely systematically evaluate image editing capabilities, or still require introducing the reference image's VAE latent to maintain identity consistency [7]. In contrast, **our model conditions solely on the instruction and the semantic features of the reference image, without using any VAE latent**. Nonetheless, we still achieve high scores on metrics that measure consistency before and after editing. This shows that

Table 1: **Image generation results on GenEval, DPG-Bench, and WISE.** † refers to methods using LLM rewriters on GenEval. Abbreviations for WISE: Cult. (Cultural), Bio. (Biology), Phy. (Physics), Chem. (Chemistry).

Models	GenEval							DPG	WISE						
	Single	Two	Count	Colors	Pos	Col-Attr	Overall	Overall	Cult.	Time	Space	Bio.	Phy.	Chem.	Overall
Generation-only Models															
SD3-Medium [17]	0.99	0.94	0.72	0.89	0.33	0.60	0.74	-	-	-	-	-	-	-	-
FLUX.1 [Dev] [29]	0.98	0.93	0.75	0.93	0.68	0.65	0.82	84.00	0.48	**0.58**	0.62	0.42	0.51	0.35	0.50
Unified Multimodal Models															
MetaQuery-XL† [45]	-	-	-	-	-	-	0.80	-	**0.56**	0.55	0.62	0.49	0.63	0.41	0.55
Tar [23]	0.99	0.92	0.83	0.85	0.80	0.65	0.84	84.19	-	-	-	-	-	-	-
BLIP3-o [6]	-	-	-	-	-	-	0.84	-	-	-	-	-	-	-	-
UniWorld-V1† [33]	0.98	0.93	0.81	0.89	0.74	0.71	0.84	-	0.53	0.55	**0.73**	0.45	0.59	0.41	0.55
OmniGen2† [63]	0.99	0.96	0.74	**0.98**	0.71	0.75	0.86	83.57	-	-	-	-	-	-	-
D-DiT [31]	0.97	0.80	0.54	0.76	0.32	0.50	0.65	-	-	-	-	-	-	-	-
Show-o [70]	0.98	0.80	0.66	0.84	0.31	0.50	0.68	-	0.28	0.40	0.48	0.30	0.46	0.30	0.35
Harmon [64]	0.99	0.86	0.66	0.85	0.74	0.48	0.76	-	0.38	0.48	0.52	0.37	0.44	0.29	0.41
MUSE-VL† [72]	-	-	-	-	-	-	0.57	-	-	-	-	-	-	-	-
Transfusion [83]	-	-	-	-	-	-	0.63	-	-	-	-	-	-	-	-
Emu3 [61]	-	-	-	-	-	-	0.66	81.60	0.34	0.45	0.48	0.41	0.45	0.27	0.39
Show-o2 [71]	**1.00**	0.87	0.58	0.92	0.52	0.62	0.76	86.14	-	-	-	-	-	-	-
Janus-Pro [10]	0.99	0.89	0.59	0.90	0.79	0.66	0.80	84.19	0.30	0.37	0.49	0.36	0.42	0.26	0.35
Mogao [32]	**1.00**	**0.97**	0.83	0.93	0.84	**0.80**	**0.89**	84.33	-	-	-	-	-	-	-
X-Omni [20]	0.98	0.95	0.75	0.91	0.71	0.68	0.83	**87.65**	-	-	-	-	-	-	-
Ming-UniVision [25]	**1.00**	0.93	0.59	0.93	**0.92**	0.70	0.85	82.12	-	-	-	-	-	-	-
BAGEL† [13]	0.98	0.95	**0.84**	0.95	0.78	0.77	0.88	85.07	0.44	0.55	0.68	0.44	0.60	0.39	0.52
UniCom	0.98	0.94	0.81	0.91	0.82	0.77	0.87	85.92	0.55	0.56	**0.73**	**0.58**	**0.66**	**0.47**	**0.58**

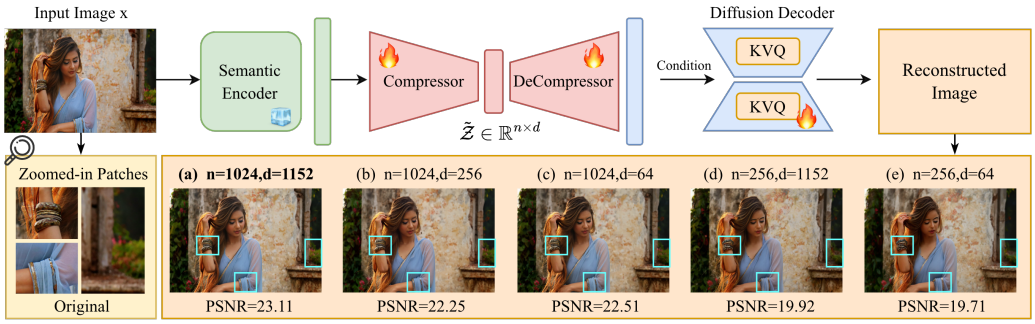


Figure 4: **Overview of the proposed diffusion decoder.** A frozen vision encoder extracts semantic features, which are compressed and decoded to condition image reconstruction. By varying n and d, we find that channel compression preserves reconstruction fidelity better than sequence reduction.

our compressed representation preserves fine-grained and structural information from the reference image. Notably, on more challenging knowledge-intensive benchmarks such as KRIS-Bench [66] and WorldEdit [59], which require complex world knowledge spanning biology, geography, and cultural understanding, our model achieves the best performance among open-source models. These results demonstrate strong capability in sophisticated editing scenarios that demand deep semantic comprehension.

Image Reconstruction. We report the quantitative evaluation on the ImageNet validation set [14] in Tab 3. Among unified tokenizers without VAE, UniCom achieves the strongest reconstruction metrics. Comparing the uncompressed (d1152) and compressed (d64) variants, reducing the channel dimension by $18\times$ incurs negligible loss in reconstruction fidelity. The qualitative superiority of our approach is further evidenced in Figures 12 to 14. As shown in the zoomed-in regions, UniCom effectively recovers high-frequency details such as small text characters and maintains consistent facial identity. Additional evaluations of small text and facial identity are provided in Appendix G. This stands in sharp contrast to prior methods relying on semantic features, which often exhibit severe degradation in these fine-grained structures.

Table 2: **Comparison of image editing capabilities.** Abbreviations for ImgEdit: Add, Adjust, Extract, Replace, Remove, Background, Style, Hybrid, Action. For GEdit-Bench, metrics include 'G-Semantic Consistency' (G-SC) and 'G-Perceptual Quality' (G-PQ). For KRIS-Bench, we report Factual (Fact.), Conceptual (Conc.), and Procedural (Proc.) knowledge scores.

Models	ImgEdit-Bench										GEdit-Bench			KRIS-Bench				WorldEdit
	Add	Adj.	Ext.	Rep.	Rm.	Bg.	Sty.	Hyb.	Act.	Overall	G-SC	G-PQ	G-Overall	Fact.	Conc.	Proc.	Overall	Overall
Generation-only Models																		
FLUX.1 Kontext [Pro] [3]	4.25	4.15	2.35	4.56	3.57	4.26	4.57	3.68	4.63	4.00	7.02	7.60	6.56	57.22	55.06	46.69	54.17	3.21
Qwen-Image [62]	4.38	4.16	**3.43**	4.66	4.14	4.38	**4.81**	3.82	**4.69**	4.27	8.00	**7.86**	**7.56**	-	-	-	-	-
Specialized Editing Models																		
Instruct-Pix2Pix [4]	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88	3.58	5.49	3.68	23.33	25.59	17.28	22.82	2.44
MagicBrush [80]	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.83	4.68	5.66	4.52	41.84	39.24	26.54	37.15	2.14
AnyEdit [76]	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45	3.18	5.82	3.21	39.26	41.88	31.74	38.55	2.09
Step1X-Edit [36]	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06	7.09	6.76	6.70	45.52	48.01	31.82	43.29	-
Unified Multimodal Models																		
OmniGen [68]	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96	5.96	5.89	5.06	33.11	28.02	23.89	28.85	2.52
Ming-Univision [25]	-	-	-	-	-	-	-	-	-	-	6.04	6.86	5.54	-	-	-	-	-
BAGEL [13]	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20	7.36	6.83	6.52	60.26	55.86	51.69	56.21	2.76
UniWorld-V1 [33]	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26	4.93	7.43	4.85	-	-	-	-	-
OmniGen2 [63]	3.57	3.06	1.77	3.74	3.20	3.57	**4.81**	2.52	4.68	3.44	7.16	6.77	6.41	57.36	44.20	47.79	49.71	2.51
TUNA [38]	**4.46**	**4.52**	2.47	**4.68**	**4.58**	**4.56**	4.73	**4.07**	**4.69**	**4.31**	7.79	7.48	7.29	-	-	-	-	-
UNICOM	4.36	4.04	3.30	4.63	4.40	4.24	4.79	3.54	**4.69**	4.22	**8.06**	7.33	7.32	**74.63**	**69.48**	**65.30**	**70.11**	**4.12**

Table 3: **Image reconstruction evaluation on the ImageNet validation set**. d denotes the feature dimension of compressed representations $\tilde{\mathbf{z}}$. For unified tokenizers without VAE, all images are resized and center-cropped to 256×256.

Tokenizer	Res.	rFID↓	PSNR↑	SSIM↑
Specialized tokenizers				
SD-VAE [18]	256	1.06	28.62	0.86
GigaTok [73]	256	0.51	21.32	0.69
VA-VAE [74]	256	0.26	28.59	0.80
DC-AE [8]	512	0.22	26.15	0.71
MAE-Tok [5]	512	0.62	-	-
TexTok [78]	512	0.73	24.45	0.66
FLUX.1[dev]-VAE [29]	512	0.06	33.65	0.93
Unified tokenizers without VAE				
UniTok [39]	256	0.38	-	-
RAE [82]	256	0.61	19.20	0.44
VILA-U [65]	256	4.23	22.24	0.61
X-Omni [8]	256	8.31	16.05	0.38
BLIP3-o [6] & EMU2 [52]	256	3.18	14.71	0.58
UNICOM (d64)	256	0.39	23.74	0.71
UNICOM (d1152)	256	**0.37**	**24.14**	**0.72**

Table 4: **Reconstruction ablation on ImageNet at** 1024×1024**.** We vary the spatial token count n, channel dimension d, and compressor architecture. Channel reduction retains the spatial grid, whereas sequence reduction decreases the token count. MLP and MHA compressors are evaluated using rFID, PSNR, and SSIM.

Setting	n	d	rFID ↓	PSNR ↑	SSIM ↑
Original	1024	1152	0.40	23.26	0.69
Dimension vs. sequence reduction					
Sequence reduction	256	1152	0.72	20.29	0.56
Dimension reduction	1024	256	**0.62**	**21.73**	**0.66**
Channel dimension with MLP compressor					
MLP, $d = 128$	1024	128	0.76	**22.21**	**0.68**
MLP, $d = 64$	1024	64	**0.55**	22.17	0.66
MLP, $d = 32$	1024	32	7.91	15.74	0.48
Channel dimension with MHA compressor					
MHA, $d = 64$	1024	64	0.56	22.61	**0.69**

4.3 Ablation Studies

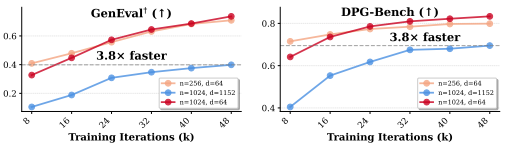
In this section, we organize ablations around three design axes required for unified multimodal modeling on reduced-scale data and simple task configurations: optimal geometry of unified semantic representations §4.3.1, semantic compressor architecture §4.3.2, and prediction pathway §4.3.3.

4.3.1 Optimal Geometry of Unified Semantic Representations

We first vary sequence length n and feature dimension d to determine the geometry of the compressed latent $\tilde{\mathbf{z}}$. The sequence length is controlled by the maximum token count, while the channel dimension is controlled by the semantic compressor.

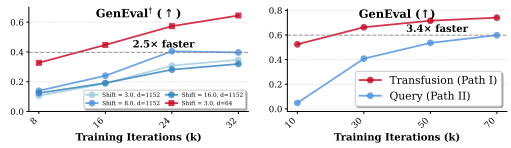
Channel compression preserves reconstruction fidelity better than sequence reduction. As shown in Tab 4 and Fig 4, reducing sequence length degrades reconstruction, while channel compression preserves most visual detail. $d = 64$–256 retains similar fidelity, whereas $d = 32$ collapses, consistent with the intrinsic dimensionality analysis in Appendix D.1: SigLIP2 features concentrate variation in a lower-dimensional subspace, so moderate channel compression removes redundancy without discarding most semantics.

Noise shifting does not remove the need for compression. Recent studies suggest shifted noise schedules for predicting uncompressed continuous visual features [82, 57]. We test this strategy on original SigLIP2 features ($n = 1024, d = 1152$). As shown in Fig 6(a), larger shifts improve the uncompressed baseline, with shift $= 8.0$ best among $\{3.0, 8.0, 16.0\}$. However, uncompressed



(a) GenEval **(b)** DPG-Bench

Figure 5: Feature dimension compression accelerates T2I training.



(a) Noise shift **(b)** Prediction path

Figure 6: Compressed latents and Transfusion-style prediction improve convergence.

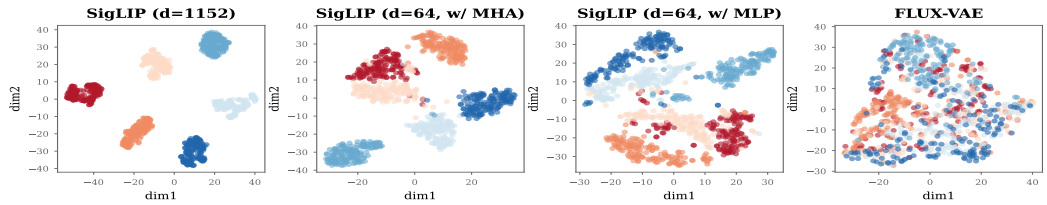


Figure 7: **T-SNE visualization of feature distributions.** We randomly select 6 distinct classes from ImageNet, sampling 150 images per class. MHA compression to 64-d preserves the structure, while MLP compression yields scattered 64-d features.

variants still converge more slowly and remain below the compressed $d = 64$ target, indicating that channel compression changes the target geometry in a way scheduling alone cannot.

Channel compressed latents improve generative efficiency. Beyond reconstruction, feature geometry strongly affects generative learning dynamics. From Fig 5, compressing to $d = 64$ converges faster and reaches higher final quality than directly predicting uncompressed features. In the 32 GPU ablation, channel compression reaches matched quality approximately $3.8\times$ faster than $d = 1152$, saving roughly 3,458 GPU hours and 18.1M training samples. At fixed $d = 64$, using $n = 1024$ rather than 256 tokens saves approximately 5.1M samples and 391 GPU hours to matched quality, despite the higher cost of processing a longer sequence. Maintaining the full token sequence ($n = 1024$) also outperforms token reduction in later training and is important for editing fidelity. We therefore select $(n = 1024, d = 64)$ as the default target.

4.3.2 Semantic Compressor Architecture

We compare a simple **MLP** compressor against the **MHA** compressor inspired by UniTok [39]. From a reconstruction perspective (Tab 4), both compressors yield comparable fidelity. However, a unified representation must also retain rich semantic structure, which we validate through feature distribution visualization and downstream visual understanding benchmarks.

Distributional Analysis (t-SNE). We visualize the feature distributions of different encoding strategies in Fig 7. As shown in Fig 7, the MHA compressor produces tightly clustered embeddings that maintain the structured distribution of the original SigLIP2 feature. In contrast, the MLP compressor yields scattered clusters with blurred semantic boundaries, indicating that attention preserves cross-token semantic context more effectively.

Table 5: Visual understanding performance on six benchmarks using compressed representations as visual inputs.

Model Config.	General VQA			Knowledge	Text-Centric	
	GQA	RealWorldQA	SEED	MMMU	ChartQA	OCRBench
Baseline	**65.25**	64.31	**74.63**	44.56	69.04	55.40
Compression Strategies						
MLP Comp.	62.80	60.39	69.92	43.00	56.80	31.70
MHA Comp.	64.01	63.14	71.75	44.11	62.12	36.00
Concatenation Strategies						
Dim. Concat	64.22	60.26	73.11	**45.00**	60.88	48.10
Seq. Concat	65.03	**64.58**	73.62	43.33	**69.24**	**55.50**

Downstream Visual Understanding Tasks. To quantify semantic capability, we employ the compressed representations $\tilde{\mathbf{z}}$ within a multimodal LLM framework [35], pre-training on LLaVA-Pretrain-557k [35] and fine-tuning on Cambrian-737k [56]. As detailed in Tab 5, the MHA compressor consistently outperforms the MLP, confirming that attention-based compression better preserves complex semantics. Although pure compression incurs minor information loss compared to the un-

compressed baseline on text-rich tasks, we find that this can be effectively mitigated by fusing the MHA-compressed features with the original SigLIP2 features. Specifically, sequence concatenation of these two representations even introduces marginal gains on fine-grained benchmarks. Thus, we adopt the MHA compressor as our final design. Appendix B reports additional understanding results for the main model, evaluates compressed features as the sole visual representation for all tasks, and analyzes the effects of understanding data scale and quality.

4.3.3 Unified Prediction Pathway

We finally ablate the prediction pathway for compressed visual latents. Besides our full-sequence Transfusion prior, we explore a query-guided variant inspired by BLIP-3o [6] (details are in Appendix E). Both variants use the same feature space and diffusion decoder. After enforcing a unified SigLIP2 space, we further add an I2I reconstruction task inspired by RECA [69] to improve structural alignment. The query-guided variant still converges more slowly than Transfusion (Fig 6(b)). Its query bottleneck compresses spatial structure before generation, while Transfusion keeps dense full-sequence correspondence for better editing fidelity and consistency (Fig 16).

5 Conclusion

In this paper, we introduce UNICOM, a unified multimodal framework that effectively integrates visual understanding and generation by compressing high-dimensional visual embeddings into a compact latent space. Based on the continuous semantic compressor, we adopt a Transfusion-style predictor to model compressed visual latents with dense spatial correspondence. Experimental results show that UNICOM not only preserves high-level semantics but also retains fine-grained visual details, achieving competitive text-to-image generation and open-source-leading performance on complex and knowledge-intensive image editing benchmarks. Furthermore, UNICOM demonstrates that compressed semantic representations can support both unified generation and understanding without relying on variational autoencoders (VAE). Future work could focus on further optimizing the compression module and expanding the application of this framework to other domains, such as video generation and multimodal reasoning.

Acknowledgments

This work is supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM113), Beijing Municipal Natural Science Foundation (Grant No. L2603008), National Natural Science Foundation of China (92470121, 62402016), National Key R&D Program of China (2024YFA1014003), Zhongguancun Academy (C20250204, C20250602), Beijing Major Science and Technology Project (Z251100008125043, Z251100008425023), and High-performance Computing Platform of Peking University.

References

[1] Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

[3] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025.

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.

[5] Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. In *Forty-second International Conference on Machine Learning*, 2025.

[6] 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*, 2025.

[7] Jiuhai Chen, Le Xue, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, An Yan, Honglu Zhou, Zeyuan Chen, Lifu Huang, et al. Blip3o-next: Next frontier of native image generation. *arXiv preprint arXiv:2510.15857*, 2025.

[8] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. *arXiv preprint arXiv:2410.10733*, 2024.

[9] Xiangyi Chen, Théophane Vallaeys, Maha Elbayad, John Nguyen, and Jakob Verbeek. Vugen: Visual understanding priors for generation. *arXiv preprint arXiv:2510.06529*, 2025.

[10] 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*, 2025.

[11] Yanzhe Chen, Huasong Zhong, Yan Li, and Zhenheng Yang. Unicode ^{2}: Cascaded large-scale codebooks for unified multimodal understanding and generation. *arXiv preprint arXiv:2506.20214*, 2025.

[12] Cheng Cui, Yubo Zhang, Ting Sun, Xueqing Wang, Hongen Liu, Manhui Lin, Yue Zhang, Tingquan Gao, Changda Zhou, Jiaxuan Liu, Zelun Zhang, Jing Zhang, Jun Zhang, and Yi Liu. Pp-ocrv5: A specialized 5m-parameter model rivaling billion-parameter vision-language models on ocr tasks, 2026. URL https://arxiv.org/abs/2603.24373.

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

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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

[16] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[17] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

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

[19] Yuan Gao, Chen Chen, Tianrong Chen, and Jiatao Gu. One layer is enough: Adapting pre-trained visual encoders for image generation. *arXiv preprint arXiv:2512.07829*, 2025.

[20] Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025.

[21] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

[22] Agrim Gupta, Linxi Fan, Surya Ganguli, and Li Fei-Fei. Metamorph: Learning universal controllers with transformers. *arXiv preprint arXiv:2203.11931*, 2022.

[23] Jiaming Han, Hao Chen, Yang Zhao, Hanyu Wang, Qi Zhao, Ziyan Yang, Hao He, Xiangyu Yue, and Lu Jiang. Vision as a dialect: Unifying visual understanding and generation via text-aligned representations. *arXiv preprint arXiv:2506.18898*, 2025.

[24] 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.

[25] Ziyuan Huang, DanDan Zheng, Cheng Zou, Rui Liu, Xiaolong Wang, Kaixiang Ji, Weilong Chai, Jianxin Sun, Libin Wang, Yongjie Lv, et al. Ming-univision: Joint image understanding and generation with a unified continuous tokenizer. *arXiv preprint arXiv:2510.06590*, 2025.

[26] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[28] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

[29] Black Forest Labs. Flux.1 model family. https://blackforestlabs.ai/announcing-black-forest-labs/, 2024.

[30] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[31] Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2779–2790, 2025.

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

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

[34] Han Lin, Jaemin Cho, Amir Zadeh, Chuan Li, and Mohit Bansal. Bifrost-1: Bridging multimodal llms and diffusion models with patch-level clip latents. *ArXiv*, abs/2508.05954, 2025. URL https://api.semanticscholar.org/CorpusID:280561829.

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[36] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

[37] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

[38] 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. *arXiv preprint arXiv:2512.02014*, 2025.

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

[40] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7739–7751, 2025.

[41] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022.

[42] Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

[43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

[44] Kaihang Pan, Wang Lin, Zhongqi Yue, Tenglong Ao, Liyu Jia, Wei Zhao, Juncheng Li, Siliang Tang, and Hanwang Zhang. Generative multimodal pretraining with discrete diffusion timestep tokens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26136–26146, 2025.

[45] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

[46] Xuerui Qiu, Yutao Cui, Guozhen Zhang, Junzhe Li, Yaqi Zhao, Xiao Zhang, Yang Li, Songtao Liu, Miles Yang, Yu Shi, Zhao Zhong, and Liefeng Bo. HYDRA: Representation harmonized tokenization for multimodal generation and understanding, 2026. URL https://arxiv.org/abs/2603.15228.

[47] 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*, pages 2545–2555, 2025.

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[50] Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024.

[51] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8802–8812, June 2021.

[52] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.

[53] 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.

[54] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

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

[56] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[57] Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2601.16208*, 2026.

[58] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

[59] Lin Wang, Feng Wang, Majun Zhang, Wentao Hu, Tao Jin, Zhou Zhao, Fei Wu, Jingyuan Chen, Sucheng Ren, and Alan Yuille. Worldedit: Towards open-world image editing with a knowledge-informed benchmark. In *ICLR*, 2026.

[60] Peiyu Wang, Yi Peng, Yimeng Gan, Liang Hu, Tianyidan Xie, Xiaokun Wang, Yichen Wei, Chuanxin Tang, Bo Zhu, Changshi Li, et al. Skywork unipic: Unified autoregressive modeling for visual understanding and generation. *arXiv preprint arXiv:2508.03320*, 2025.

[61] 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.

[62] 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*, 2025.

[63] 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*, 2025.

[64] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. *arXiv preprint arXiv:2503.21979*, 2025.

[65] 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. *arXiv preprint arXiv:2409.04429*, 2024.

[66] Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. Kris-bench: Benchmarking next-level intelligent image editing models. *arXiv preprint arXiv:2505.16707*, 2025.

[67] xAI. Grok-1.5 vision preview, 2024. URL https://x.ai/news/grok-1.5v. Company news post.

[68] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13294–13304, 2025.

[69] Ji Xie, Trevor Darrell, Luke Zettlemoyer, and XuDong Wang. Reconstruction alignment improves unified multimodal models. *arXiv preprint arXiv:2509.07295*, 2025.

[70] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

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

[72] Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24135–24146, 2025.

[73] Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. *arXiv preprint arXiv:2504.08736*, 2025.

[74] 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*, pages 15703–15712, 2025.

[75] 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.

[76] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26125–26135, 2025.

[77] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[78] Kaiwen Zha, Lijun Yu, Alireza Fathi, David A Ross, Cordelia Schmid, Dina Katabi, and Xiuye Gu. Language-guided image tokenization for generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15713–15722, 2025.

[79] Guozhen Zhang, Xuerui Qiu, Yutao Cui, Tianhui Song, Changlin Li, Junzhe Li, Tao Huang, Xiao Zhang, Yang Li, Jianbing Wu, Miles Yang, Zhao Zhong, Liefeng Bo, and Limin Wang. HYDRA-X: Native unified multimodal models with holistic visual tokenizers, 2026. URL https://arxiv.org/abs/2606.13289.

[80] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS*, 2023.

[81] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

[83] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

A Implementation details

A.1 Decoder Setup.

As illustrated in Figure 4, we employ the SigLIP2-SO400M-Patch16-NaFlex [58] as our vision encoder. To enable high-fidelity image generation, we initialize the diffusion backbone with FLUX.1-dev [29]. Crucially, to bridge the dimensionality gap between dense semantic features and the generative latent space, we implement a Multi-Head Attention (MHA) semantic compressor following UniTok [39], which effectively preserves rich semantics during feature compression. We optimize both the compression module and the diffusion model on a high-quality internal dataset. We adopt a robust multi-resolution training strategy. Specifically, we define a grid of 33 discrete aspect ratio buckets, ranging from 1:4 (vertical) to 4:1 (horizontal), anchored at a base resolution of 1024×1024. Training is conducted with a global batch size of 256. We observe rapid convergence, achieving high-quality reconstruction performance within 10K steps. However, we extend training to a total of 50K steps to ensure the full recovery of fine-grained visual details.

A.2 Unified Model Setup.

We initialize the language component with Qwen2.5-7B-Instruct [55] and the vision encoder with SigLIP2-SO400M-Patch16-NaFlex [58]. Visual features are mapped into the LLM's embedding space via a two-layer MLP connector. As outlined in Table 15, we employ a multi-stage training pipeline using a dynamic mixture of curated data. Throughout all phases, we keep the SigLIP2 vision encoder frozen and adopt the same multi-resolution strategy utilized in the decoder setup, processing images with dynamic token counts up to a maximum of 1,024. The training process proceeds as follows: (1) **Alignment**: We establish fundamental visual understanding capabilities by optimizing both the MLP connector and the language model parameters. (2) **Pre-training**: Subsequently, we inject generative capabilities using a massive corpus of text, image-text pairs, and image-text-image sequences. (3) **Continued Training**: To enhance performance on complex tasks such as inpainting and subject-driven generation, we strategically increase the sampling ratio of image-text-image pairs. (4) **Supervised Fine-tuning**: Finally, the model undergoes fine-tuning on high-quality curated data to ensure precise instruction following across all modalities.

A.3 Ablation Setup.

For efficient ablation analysis in Section 4.3, we switch to reduced data scales and simple task configurations. When investigating reconstruction fidelity, we train the compression and diffusion modules on a subset of high-quality internal images. Regarding generative modeling dynamics, we train the model on a filtered text-to-image dataset to assess convergence speed and generation quality. We evaluate the model's multimodal understanding capabilities on multiple benchmarks, including general VQA benchmarks such as GQA [26], RealWorldQA [67], and SEED-Bench [30]; knowledge-intensive benchmarks such as MMMU [77]; and text-centric benchmarks including ChartQA [41], OCRBench [37].

B Multimodal Understanding

We report understanding results for the unified model and examine how the scale and quality of supervision affect the use of compressed visual features.

B.1 Understanding Performance of the Unified Model

A single checkpoint of our main model supports understanding, generation, and editing. It uses original SigLIP2 features for understanding and compressed features for generation and editing. Table 6 reports its understanding performance alongside BAGEL, complementing the generation and editing results in the main text. Both models use Qwen-2.5-7B and SigLIP2, although their training data and optimization differ.

Table 6: Multimodal understanding of the final unified model. Higher is better for all metrics.

Model	MME	MMBench-en	MMMU	MMVP	OCRBench
BAGEL (7B)	2388	85.0	55.3	69.3	73.3
UNICOM (7B)	2268	82.5	51.0	60.7	72.0

Further training improves understanding and generation together. From pre-training to continued training and supervised fine-tuning, MMVP increases from 55.26 to 60.67, MMMU from 46.22 to 51.04, and OCRBench from 66.2 to 72.0. GenEval increases from 0.77 to 0.87 and DPG-Bench from 82.60 to 85.92 over the same stages. These gains reflect the combined effects of additional training, data, and instruction tuning.

B.2 Understanding Data Scale and Quality

Table 7 compares five configurations of visual representations and understanding supervision. The enriched corpus replaces part of the original 36M samples with higher quality document, chart, and scene text OCR data, keeping the total size unchanged. The large-scale results are evaluated after pre-training, with the compressor frozen in the variant that uses compressed features for all tasks.

Table 7: Text-centric understanding under different representations and supervision. Z: original features; $\tilde{z}$: compressed features.

Run	Visual input	Understanding data	OCRBench	ChartQA
A	Z	1.3M	55.4	69.04
D	$\tilde{z}$	1.3M	36.0	62.12
B	Z	36M	66.2	73.72
C	Z	Enriched 36M	69.2	74.88
E	$\tilde{z}$	Enriched 36M	63.4	69.36

More and better supervision substantially improves understanding from compressed features, raising OCRBench from 36.0 to 63.4. Under matched enriched supervision, the OCRBench and ChartQA gaps to original features narrow to 5.80 and 5.52 points, respectively. These results demonstrate substantial recovery while leaving a measurable difference between the two representations.

C Architecture Ablations

We examine the transfer of the compressed semantic interface across language backbones, visual encoders, and image decoders. These comparisons complement the controlled representation ablations in the main text.

C.1 Language and Generative Backbones

We compare two complete systems trained with the same compressed semantic targets and SigLIP2 encoder, replacing both the language backbone and generative decoder (Table 8). HunyuanVideo [28] serves as the alternative generative decoder.

Table 8: Complete backbone configurations with SigLIP2 and compressed semantic targets.

LLM / decoder	GenEval	DPG	GEdit	MMBench	MMVP
Qwen-2.5-7B / FLUX.1-dev	0.87	85.9	7.3	82.5	60.7
Hunyuan-LLM / HunyuanVideo	0.89	84.9	6.5	76.7	42.1

Both configurations train stably, demonstrating that the compression recipe transfers across backbone configurations. Both achieve competitive generation, while the Qwen-2.5-7B and FLUX.1-dev configuration performs better on understanding and editing. In contrast, raw SigLIP2 targets ($d = 1152$) converge slowly, and raw Qwen-2.5-VL ViT targets ($d = 3584$) show no sign of convergence within the matched budget. Together with the controlled ablations in the main text, these results support compression as an effective means of simplifying semantic prediction. Since both backbones change together, this comparison does not isolate their individual contributions.

C.2 Visual Encoders and Generative Decoders

We compare encoder and decoder choices using uncompressed features on ImageNet validation at 256×256 (Table 9). We vary the encoder with FLUX.1-dev fixed and vary the decoder with SigLIP2 fixed. The alternative encoder is the vision tower of Qwen2.5-VL [2].

Table 9: Encoder and diffusion-decoder comparisons using original semantic features.

Encoder	Decoder	PSNR ↑	SSIM ↑
SigLIP2 ($d = 1152$)	FLUX.1-dev	24.14	0.72
Qwen-2.5-VL ViT ($d = 3584$)	FLUX.1-dev	23.79	0.69
SigLIP2 ($d = 1152$)	HunyuanVideo	23.64	0.69

SigLIP2 outperforms Qwen-2.5-VL ViT with the same decoder, and FLUX.1-dev outperforms HunyuanVideo with the same encoder. These results motivate SigLIP2 and FLUX.1-dev as our default configuration. We did not evaluate DINOv2 [43] because adapting an encoder without language alignment to the same unified setting would require additional data and compute beyond our budget. A matched comparison remains future work.

C.3 Lightweight Reconstruction Decoders

We use a lightweight deterministic decoder to test whether compressed features support reconstruction without a powerful diffusion prior. Trained from scratch, the decoder combines a token Transformer with eight layers and sixteen heads, convolutional upsampling across multiple scales, direct patch prediction, and detail refinement at high resolution. Training uses MSE, L1, and LPIPS [81] losses at 512×512, and evaluation is performed at 256×256 (Table 10). The decoder contains 141.5M parameters, approximately $85\times$ fewer than FLUX.1-dev. Including the 13.7M-parameter MHA compressor and decompressor brings the compressed configuration to 155.2M parameters.

With the deterministic decoder, $18\times$ channel compression reduces PSNR by 0.35 dB and SSIM by 0.01. Accurate reconstruction is therefore attainable without the FLUX diffusion prior. Compared with FLUX, the deterministic decoder achieves higher PSNR and SSIM but worse rFID, while FLUX

Table 10: Decoder capacity and compression on ImageNet validation at 256×256. Parameter counts for the lightweight compressed setting include compression and decompression.

Decoder	Feature	Parameters	PSNR ↑	SSIM ↑	rFID ↓
FLUX.1-dev	Original ($d = 1152$)	12B	24.14	0.72	0.37
FLUX.1-dev	Compressed ($d = 64$)	12B	23.74	0.71	0.39
Deterministic	Original ($d = 1152$)	141.5M	25.25	0.75	1.63
Deterministic	Compressed ($d = 64$)	155.2M	24.90	0.74	1.74

produces sharper facial features, small text, and textures in qualitative comparisons. The difference reflects both decoder capacity and training objectives, rather than capacity alone.

D Channel Compression Preserves Information Better than Sequence Compression

We analyze why sequence reduction weakens reconstruction and generation compared with channel compression. The key distinction is where information is removed: channel compression acts on redundant feature directions, while sequence compression merges spatial tokens that each correspond to distinct image regions. We therefore first quantify channel redundancy, then show that sequence reduction loses spatial correspondence, and finally connect this information loss to reconstruction and T2I training dynamics.

D.1 Intrinsic Dimensionality of SigLIP2 Features

To quantify channel redundancy, we estimate the intrinsic dimensionality of SigLIP2 features using the participation ratio (PR) of the feature covariance spectrum. Given covariance eigenvalues $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_M$, PR is defined as

$$\mathrm{PR} = \frac{\left(\sum_i \mu_i\right)^2}{\sum_i \mu_i^2}. \tag{7}$$

Table 11: Channel dimension sweep.

Dim. (d)	**PSNR ↑**	**SSIM ↑**
256	21.73	0.66
128	22.21	0.68
64	22.17	0.66
32	15.74	0.48

PR equals 1 when all variance lies in one direction and equals M when variance is uniformly distributed across all dimensions. On SigLIP2 features with $M = 1152$ extracted from 1K ImageNet validation images, we obtain $\mathrm{PR} \approx 114$. This indicates that visual-language pre-training concentrates most variation on a lower-dimensional subspace, so moderate channel compression removes redundant directions rather than discarding most semantic content. Table 11 further shows that reconstruction quality is stable from $d = 64$ to $d = 256$, while $d = 32$ loses substantial information. The best PSNR appears at $d = 128$, which is closest to the estimated PR. We choose $d = 64$ because it remains comparable in reconstruction fidelity while also producing a compact prediction target that accelerates T2I training, as shown in Fig 5.

D.2 Sequence Reduction Loses Spatial Correspondence

Compression operators. Let an image of resolution $H \times W$ be encoded into a patch grid with $n = HW/P^2$ tokens and feature dimension d, where P is the patch size. Channel compression preserves the token grid and reduces only the feature dimension,

$$\mathcal{C}_{\mathrm{ch}} : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d'}, \quad d' \ll d. \tag{8}$$

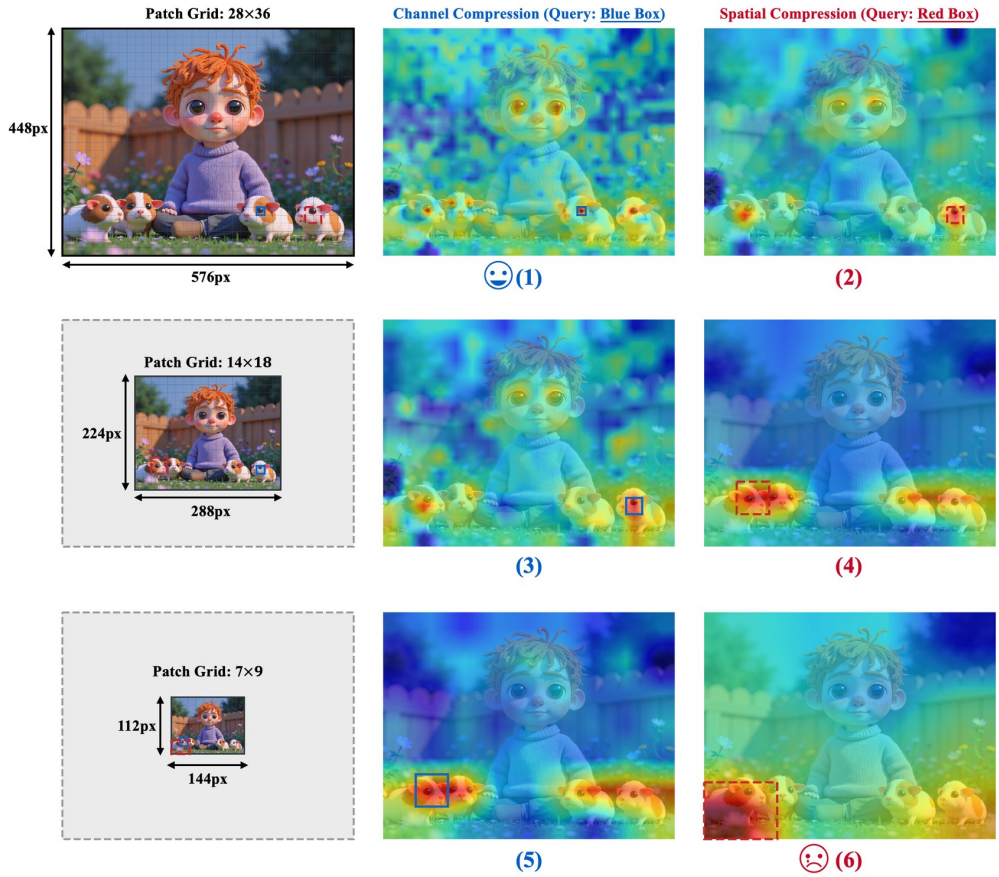


Figure 8: **Patch-level cosine similarity under different resolutions.** We select the query patch marked by a box and compare similarity heatmaps across three resolutions. Channel compression preserves localized semantic correspondence, while spatial compression exhibits similarity bleeding as resolution decreases.

Spatial compression instead reduces the number of tokens,

$$\mathcal{C}_{\text{sp}} : \mathbb{R}^{n\times d} \to \mathbb{R}^{(n/r)\times d}, \tag{9}$$

where each output token aggregates information from roughly r input patches. Since each patch covers a $1/n$ fraction of the image grid, a spatially compressed token covers about r/n of the image. When the resolution decreases, n becomes smaller and r/n grows, so each merged token spans a larger visual region. This makes spatial compression increasingly resolution-dependent, while channel compression remains tied to the original spatial sampling.

Patch correspondence under resolution changes. Figure 8 visualizes this behavior with patch-level cosine similarity heatmaps. Channel-compressed features preserve sharp semantic correspondence across resolutions, such as localizing semantically matched object parts. Spatial compression increasingly spreads similarity over unrelated regions at lower resolutions, indicating that merging spatial tokens can destroy correspondence that is difficult to recover later.

D.3 Dimensionality Compression Improves Reconstruction and Training

Figure 9 further links the spatial information loss above to reconstruction behavior: sequence reduction converges quickly at first but saturates at lower PSNR because fewer tokens provide insufficient spatial evidence for recovering fine details. Channel compression has a harder early optimization stage, but it keeps the full $n = 1024$ spatial grid and eventually approaches the uncompressed baseline. For generation, the same representation shape is also easier to predict. Fig 5 shows that feature

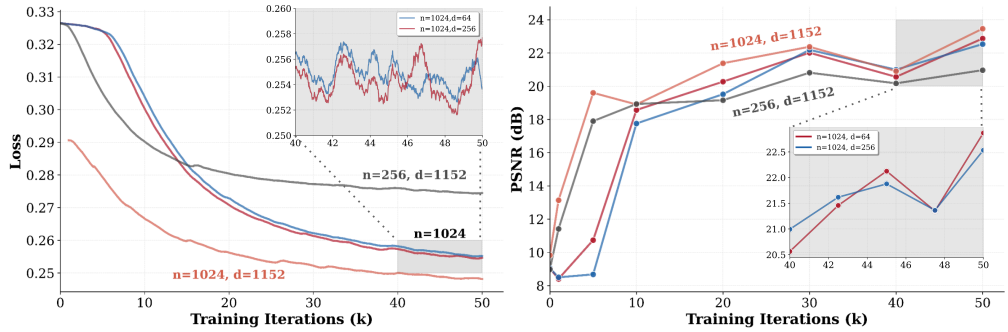


Figure 9: **Reducing dimensions maintains higher reconstruction fidelity than sequence reduction.** We compare the reconstruction training dynamics (Loss and PSNR) of different compressed representations ($\tilde{\mathbf{z}}$).

dimension compression accelerates T2I training and reaches stronger final quality than directly predicting high-dimensional features.

Summary. Together, these analyses explain why channel compression is the preferred representation shape for UniCom. It exploits the low intrinsic dimensionality of semantic channels, preserves the dense spatial token grid required for alignment and editing, and provides a lower-dimensional target that is easier for the generative model to predict, as evidenced by the T2I training curves in Fig 5. Sequence compression reduces token count, but it removes spatial evidence that is unique to each patch and becomes less robust as image resolution decreases, leading to weaker reconstruction fidelity.

E Query-Guided Prediction Alternative

Besides the Transfusion prior used in the main model, we also examine a query-guided prediction architecture as an alternative design. This variant decouples the understanding and generation roles: a frozen multimodal LLM provides semantic conditioning, while a small trainable connector maps this conditioning signal into the compressed visual latent space used by the diffusion decoder.

Concretely, given a text condition $\mathbf{c}$, we concatenate the text tokens with a set of learnable **Meta-Queries** $\mathcal{Q} \in \mathbb{R}^{M \times d}$, where M is the number of queries and d matches the hidden dimension of the multimodal LLM. The frozen model processes $[\mathbf{c}; \mathcal{Q}]$, and we extract the output states corresponding to the query positions as the conditional signal $\mathbf{h}_{\mathcal{Q}} \in \mathbb{R}^{M \times d}$. A lightweight trainable connector then maps $\mathbf{h}_{\mathcal{Q}}$ to the input space of the flow-matching decoder.

This design is useful as a controlled comparison because it uses the same semantic compressor and diffusion decoder as UniCom. However, it compresses the visual condition into a small number of query states before generation. As discussed in Section 4.3.3, this bottleneck slows convergence and weakens spatial correspondence compared with the Transfusion prior, which directly models the full mixed sequence of text tokens and compressed visual latents.

F Additional Analysis

We provide two additional studies on the semantic compressor and the role of visual-language initialization.

F.1 Scaling Behavior of Compressor Architectures

We examine this question from two axes: training data scale and model size.

Training data scaling. We compare the MHA compressor trained on 7M samples with MLP compressors trained on 7M and 12M samples. KNN evaluates local semantic preservation by checking

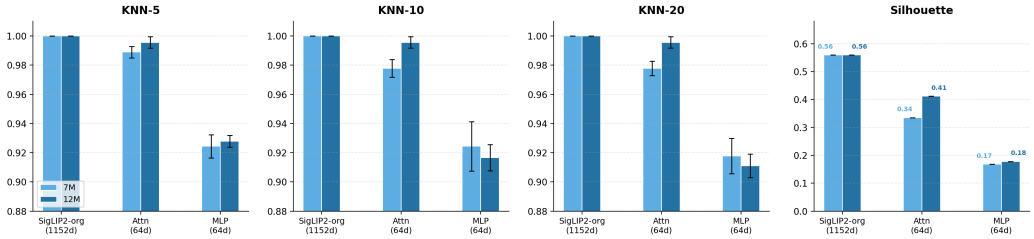


Figure 10: **Semantic structure of compressed features under data scaling.** KNN accuracy evaluates whether local neighborhoods in the compressed space preserve semantic labels, while Silhouette measures cluster separation. The MHA compressor preserves stronger neighborhood and cluster structure than the MLP compressor at both data scales.

Table 12: **Data scaling of compressor architectures.**

Benchmark	MHA-7M	MLP-7M	MLP-12M
GQA	63.83	60.92	61.97
RealWorldQA	64.58	57.65	57.52
SEEDBench	71.78	65.73	67.01
MMMU	44.00	41.44	42.33
ChartQA	63.24	22.08	46.68
OCRBench	40.50	10.90	22.40

whether nearest neighbors in the compressed feature space retain consistent labels, while the Silhouette score measures the separation of semantic clusters. As shown in Figure 10, the MHA compressor maintains stronger local neighborhoods and cluster separation than the MLP compressor. The downstream results in Table 12 show the same trend: although increasing MLP data from 7M to 12M improves some text-centric tasks, MLP-12M still underperforms MHA-7M on all benchmarks, including GQA, RealWorldQA, SEEDBench, MMMU, ChartQA, and OCRBench. This indicates that the semantic gap is not only a matter of data scale.

Model size scaling. The default MLP compressor has about 1.2M parameters, while the MHA compressor has about 13.6M parameters. Scaling the MLP to a deeper bottleneck with roughly 5M parameters does not close the gap and fails to converge reliably in reconstruction training, with PSNR below 10 at 10k steps compared with 18.57 for the base MLP. This behavior reflects a structural limitation: an MLP transforms each spatial token independently, whereas the MHA compressor explicitly models cross-token relations. We therefore adopt the MHA compressor as the default architecture for preserving semantic structure under channel compression.

F.2 Visual Understanding Informs Image Generation

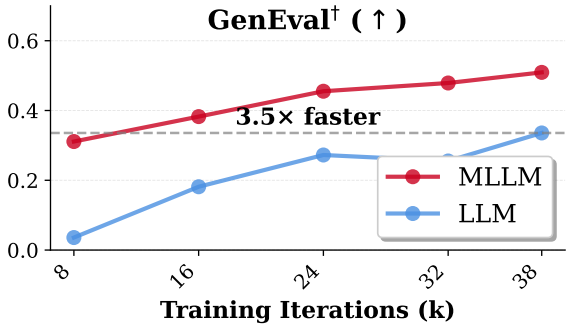


Figure 11: **Visual understanding initialization improves generation.**



Figure 12: **Visual comparison of image reconstruction results**. Notably, our method (d64) maintains high fidelity in high-frequency details (e.g., text characters) and preserves facial identity better than these semantic-based baselines, achieving quality comparable to the specialized Flux VAE.

To investigate whether semantic knowledge acquired from understanding tasks contributes to generative performance, we compare two initialization strategies for unified training: (1) initializing from the language model Qwen-2.5-7B [55], and (2) initializing from a pre-trained VLM, where Qwen-2.5-7B coupled with SigLIP2 features has already been trained on image-to-text understanding tasks. Both models are subsequently trained on the same unified generation task under identical settings.

As shown in Figure 11, the VLM-initialized model converges faster and reaches higher GenEval scores than the language-only initialization. This suggests that dense visual-language alignment learned during understanding provides a useful prior for generation. The result further supports our choice of a shared semantic representation rather than separate understanding and generation feature spaces.

G More Reconstruction Results

G.1 Qualitative Comparisons

We provide comprehensive qualitative evaluations in Figures 12 to 14. The first two figures compare UniCom with semantic-representation baselines and the Flux VAE, showing that channel-compressed features preserve high-frequency details such as text, small objects, and facial identity. Figure 14 further provides direct side-by-side comparisons between our channel-compressed reconstructions (d64) and the original images, demonstrating that the compressed representation retains textures, lighting gradients, and clothing lettering across diverse scenes.

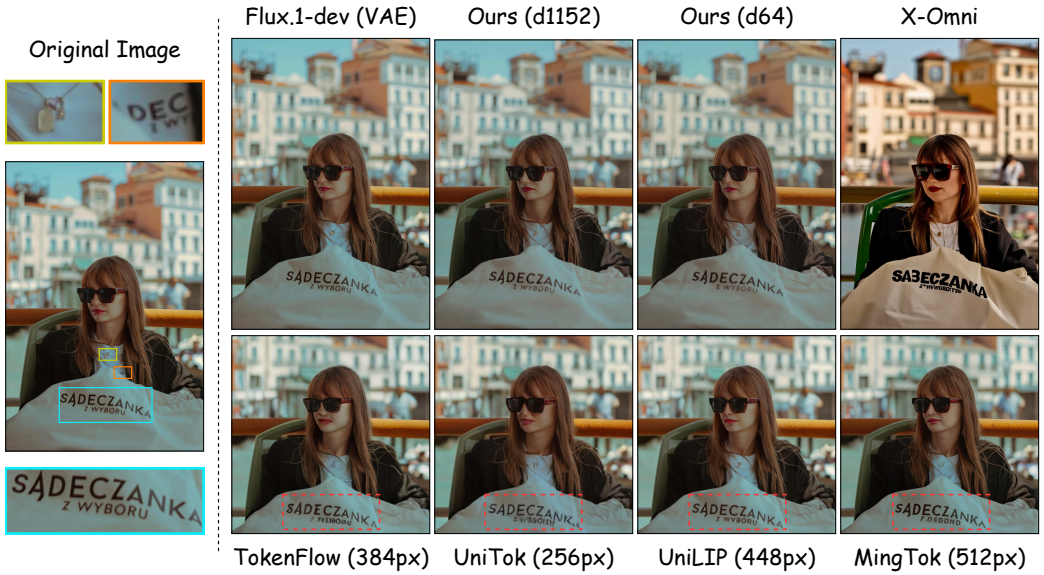


Figure 13: Additional visual comparison of image reconstruction results.



Figure 14: **Qualitative reconstruction samples using Decoder (d64). Left:** Reconstructed image. **Right:** Original image. Even with an 18× channel compression ratio, UNICOM accurately recovers complex high-frequency details.

G.2 Quantitative Evaluation of Text and Identity

We compare original ($d = 1152$) and compressed ($d = 64$) features on small text reconstruction and facial identity preservation.

Small-text reconstruction. We evaluate 497 TextOCR [51] validation images with approximately 18K word boxes. Text heights are measured after the model's resizing and cropping to the 1024-pixel input. Recognition-only PP-OCRv5 [12] reads crops from original images and reconstructions, with exact-match accuracy and $1 - \mathrm{CER}$ reported in Table 13. CER denotes character error rate. Reconstructions use decoders trained with original or compressed features.

Table 13: TextOCR recognition on original crops and reconstructed crops. Each entry is exact-match accuracy / $1 - \mathrm{CER}$; higher is better.

Text height at input	Original image	Original feature ($d = 1152$)	Compressed ($d = 64$)
> 40 px	0.292 / 0.528	0.255 / 0.478	0.254 / 0.470
20–40 px	0.395 / 0.717	0.212 / 0.460	0.194 / 0.435
10–20 px	0.451 / 0.790	0.078 / 0.198	0.069 / 0.186
< 10 px	0.440 / 0.664	0.014 / 0.055	0.009 / 0.051

The recognition gap between compressed and original features is at most 0.025. Both reconstructions degrade sharply below 10 pixels, indicating that compression adds relatively little error beyond the limitations of the semantic reconstruction pipeline.

Facial identity. We evaluate 1,000 FFHQ [27] images using ArcFace [15], defining identity retention as cosine similarity greater than 0.4. Faces are detected in every image for all methods. Table 14 reports 99.7% retention with compressed features, compared with 99.8% for original features and 99.9% for FLUX VAE. The lower mean similarity indicates a remaining fidelity gap.

Table 14: Facial identity preservation on 1,000 FFHQ images.

Reconstruction	Mean cosine similarity ↑	Identity retention ↑
FLUX VAE	0.993	0.999
Original semantic feature ($d = 1152$)	0.905	0.998
Compressed semantic feature ($d = 64$)	0.889	0.997

Why text reconstruction and OCR understanding differ. Reconstructing character appearance does not require the same symbolic readout as OCR understanding. With the compressor frozen, additional understanding supervision improves this readout (Appendix B.2) without changing the representation or its reconstruction capability.

H More Text-to-Image Generation Results

We provide additional text-to-image examples in Figure 15 to illustrate prompt-following ability across diverse instructions. The prompts cover both short descriptions and longer compositional requests, including portraits, animals, object-centric scenes, and multi-entity layouts.

I More Image Editing Results

We present comprehensive qualitative results to demonstrate the versatility of UniCom across diverse image editing and generation scenarios, spanning single-image editing (Figure 17), multi-element composition (Figure 18), and knowledge-driven intelligent editing (Figure 19).

Single Image Editing. Figure 17 illustrates the broad editing capabilities of UniCom across six representative tasks: (1) *Add/Remove/Extract*, which covers adding objects, removing elements, and extracting subjects from scenes; (2) *Replace*, which substitutes objects or materials while faithfully

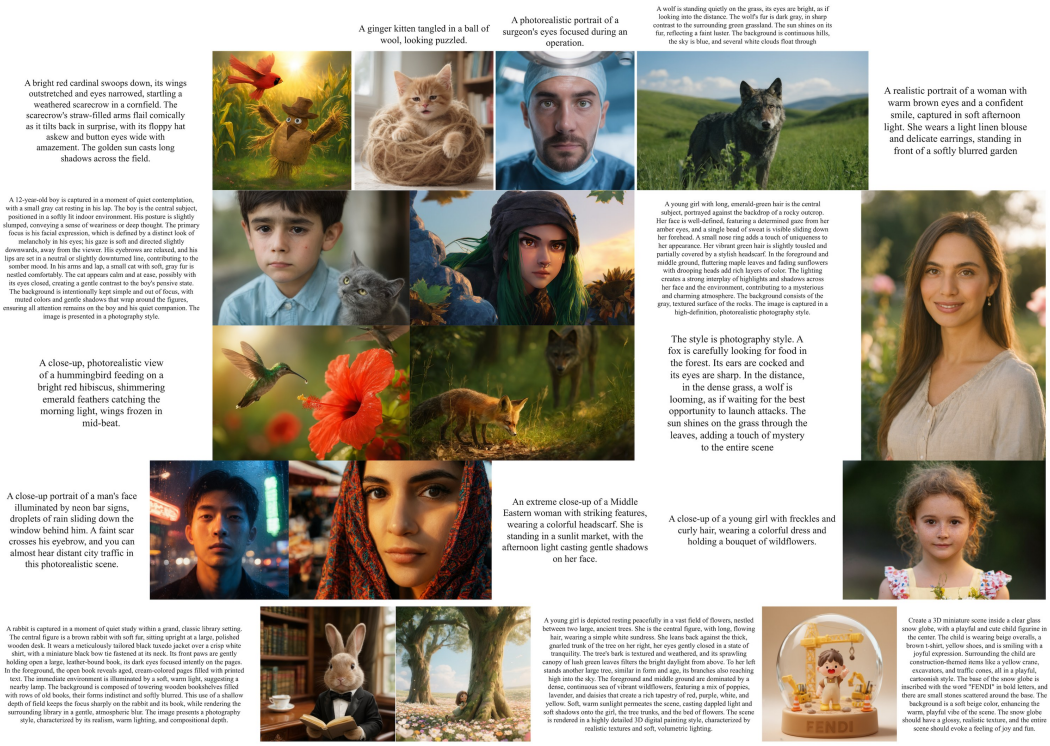


Figure 15: **Additional text-to-image generation results.** UNICOM follows diverse prompts with varied length and semantics, preserving object identity, scene layout, style, and fine-grained visual attributes across open-ended generation cases.

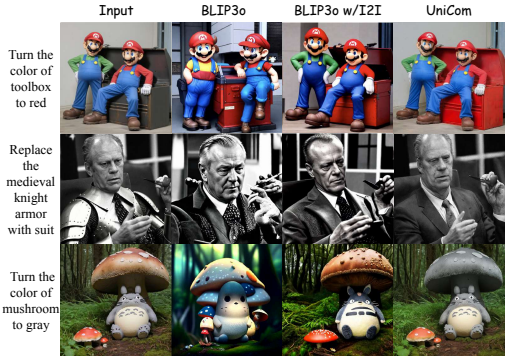


Figure 16: Our unified prediction architecture (UNICOM) preserves finer spatial layout and structural consistency, while query-based methods struggle with spatial correspondence and editing fidelity.

Table 15: **Training details and computational cost of UNICOM.**

	Alignment	PT	CT	SFT
Hyperparameters				
Learning rate	2.0×10^{-5}	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-5}
LR scheduler	Cosine	Constant	Constant	Constant
Weight decay	0.0	0.01	0.01	0.01
Gradient norm clip	1.0	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 10^{-6}$)			
Loss weight ($\lambda_{\text{diff}} : \lambda_{\text{text}}$)	-	5:1	5:1	5:1
Warm-up steps	600	5000	5000	100
Training steps	20K	115K	60K	7K
Data sampling ratio				
Text-only	0.2	0.25	0.20	0.10
Image-Text (T2I)	0.0	0.45	0.40	0.20
Image-Text (I2T)	0.8	0.15	0.20	0.40
Image-Text-Image	0.0	0.15	0.20	0.30
Computational cost				
Hardware	128 H20	256 H20	512 H20	512 H20
Time cost	∼18h	∼100h	∼100h	∼10h

preserving contextual layout; (3) *Style Transfer*, which transforms image appearance into various artistic styles; (4) *Background*, which modifies or replaces image backgrounds; (5) *Subject-Driven Generation*, which generates new images conditioned on reference subjects; and (6) *Controllable Generation*, which produces images guided by spatial or structural control signals. These results confirm that our compressed continuous semantic representation retains sufficient fine-grained information to enable precise, instruction-following edits.

Multi-Element Composition. Figure 18 demonstrates the ability of UNICOM to compose multiple reference subjects into a coherent scene. Despite the challenge of maintaining the identity of each element while harmonizing spatial arrangement and stylistic consistency, our model produces visually plausible compositions. This capability benefits from the rich semantic priors preserved by the channel-compressed representation, enabling faithful identity preservation even without relying on VAE-based reconstruction.

Intelligent Image Editing. Beyond conventional instruction-based editing, we further evaluate UNICOM on benchmarks that require world knowledge and contextual reasoning, namely WorldEdit [59] and Kris [66]. Unlike systems that rely on explicit editing commands, these benchmarks require models to integrate world knowledge priors and perform reasoning over semantic representations before generating the desired output. Figure 19 showcases representative examples that involve biological transformations, geographical understanding, and cultural context. Rather than following literal instructions, the model must infer what transformation is semantically appropriate given the input, which highlights the advantage of unifying understanding and generation within a shared semantic space.

Quantitatively, as shown in Table 16, UNICOM achieves a high overall score of 4.35/5 on WorldEdit-Test, demonstrating strong performance across diverse cause categories. These results underscore the effectiveness of semantic representation for generation: by leveraging rich visual-language priors, UNICOM produces more coherent and contextually accurate edits, outperforming most traditional methods that rely solely on explicit instructions. The integration of world knowledge and reasoning significantly enhances the model's ability to handle complex and context-sensitive editing tasks.

Table 16: The performance of both commercial and open-source models on image editing tasks in WorldEdit-Test, evaluated across different causes and metrics.

Cause Category	Metric	GPT-4o	Nano-Banana	Seedream4.0	UNICOM	WorldEdit	Flux-Kontext	Bagel-Think	Bagel	Omnigen	Omnigen2	Emu2	Anyedit	Ip2p	Magicbrush
Time	VC	4.02	4.22	3.92	3.08	4.18	4.6	2.38	2.34	3.02	2.22	1.48	1.63	3.04	1.40
	VQ	5.00	5.00	4.86	4.78	4.55	4.88	4.52	4.02	4.20	4.35	4.56	3.25	4.51	4.28
	IF	3.86	3.90	4.25	4.08	3.71	1.46	2.46	1.66	1.38	1.33	1.54	1.79	1.51	1.19
	KP	4.20	4.14	4.41	4.38	3.94	1.42	2.86	1.80	1.48	1.25	1.75	1.92	1.65	1.34
	Avg	4.27	4.32	4.36	4.08	4.10	3.09	3.06	2.46	2.52	2.29	2.33	2.15	2.68	2.05
Temperature	VC	4.02	4.22	4.06	3.40	3.52	3.86	2.08	1.94	2.54	2.26	1.38	1.45	2.62	1.81
	VQ	5.00	4.96	4.78	4.34	4.36	4.54	4.06	4.04	4.00	4.46	4.65	2.92	4.56	4.28
	IF	4.54	4.20	4.50	4.38	4.04	2.08	3.52	1.80	1.54	1.46	1.29	1.51	1.32	1.28
	KP	4.66	4.26	4.48	4.20	3.90	1.88	3.76	1.90	1.38	1.37	1.38	1.45	1.44	1.30
	Avg	4.56	4.41	4.46	4.06	3.96	3.09	3.36	2.42	2.37	2.39	2.17	1.83	2.49	2.17
Humidity	VC	4.50	4.40	4.40	3.66	3.82	4.40	2.68	2.32	2.28	1.94	1.76	1.62	2.86	1.40
	VQ	4.94	4.90	4.94	4.62	4.58	4.94	4.42	4.32	4.14	4.50	4.56	3.52	4.62	4.28
	IF	3.92	3.52	4.74	4.56	3.28	1.52	3.36	2.22	1.34	1.54	1.88	1.90	1.72	1.36
	KP	4.26	3.76	4.78	4.22	3.44	1.60	3.36	2.16	1.44	1.64	2.20	2.06	1.90	1.40
	Avg	4.41	4.15	4.72	4.20	3.78	3.12	3.46	2.76	2.30	2.41	2.60	2.28	2.78	2.11
Acidity	VC	4.46	4.36	3.92	3.42	3.94	4.74	2.74	2.60	3.14	1.88	1.34	1.38	2.74	1.60
	VQ	5.00	4.76	4.94	4.75	4.76	4.88	4.28	4.40	4.28	4.68	4.70	3.32	4.40	4.26
	IF	3.82	3.58	4.38	4.02	3.50	1.20	2.36	1.40	1.10	1.08	1.12	1.22	1.18	1.12
	KP	3.70	3.44	4.06	3.91	3.68	1.20	2.34	1.40	1.06	1.08	1.14	1.26	1.14	1.04
	Avg	4.25	4.04	4.33	4.03	3.97	3.01	2.93	2.45	2.40	2.18	2.08	1.80	2.37	2.01
Light	VC	3.70	4.40	4.24	3.34	3.82	4.64	2.84	2.36	2.44	1.88	1.45	1.43	2.66	1.54
	VQ	4.78	4.92	4.96	4.51	4.68	4.84	4.46	4.26	4.26	4.48	4.71	2.98	4.50	4.20
	IF	4.40	2.88	4.44	3.77	3.74	1.86	3.60	2.64	1.98	2.31	2.22	2.11	1.62	1.41
	KP	4.64	3.50	4.58	4.46	4.12	2.22	3.84	3.00	2.20	2.67	2.67	2.28	2.06	1.83
	Avg	4.38	3.93	4.56	4.02	4.09	3.39	3.69	3.07	2.72	2.83	2.77	2.20	2.71	2.25
Break	VC	4.80	4.90	4.64	4.14	4.74	4.88	4.00	2.98	2.98	1.80	1.36	1.46	2.36	1.40
	VQ	4.92	4.78	4.94	4.82	4.48	4.72	4.12	4.06	4.54	4.46	4.64	3.28	4.50	4.14
	IF	4.06	4.34	4.58	4.66	4.56	2.36	3.10	2.18	1.40	1.10	1.36	1.62	1.08	1.22
	KP	3.84	4.14	4.38	4.61	4.36	2.06	2.72	1.92	1.24	1.02	1.30	1.46	1.10	1.14
	Avg	4.41	4.54	4.64	4.56	4.54	3.51	3.49	2.79	2.54	2.10	2.17	2.00	2.26	1.98
Inflate	VC	4.53	4.88	4.63	3.78	4.20	4.54	3.00	2.53	2.38	1.94	1.59	1.31	2.58	1.56
	VQ	4.94	4.98	4.65	4.58	4.46	4.92	4.46	4.12	3.98	4.47	4.57	3.20	4.70	4.15
	IF	4.22	3.56	3.82	4.56	3.96	1.48	3.34	3.14	2.12	2.67	2.33	2.08	1.64	1.65
	KP	4.31	3.36	3.51	3.88	3.70	1.46	3.06	2.94	1.78	2.37	2.37	1.74	1.5	1.60
	Avg	4.50	4.20	4.15	4.20	4.08	3.10	3.47	3.18	2.57	2.86	2.71	2.08	2.61	2.24
Squeeze	VC	4.44	4.80	4.56	3.52	4.14	4.68	3.12	2.46	2.92	2.26	1.53	1.48	2.36	1.79
	VQ	4.94	4.94	4.76	4.77	4.34	4.82	3.88	4.12	4.24	4.62	4.51	3.22	4.24	4.27
	IF	2.86	2.54	3.78	2.59	3.38	1.22	2.56	1.90	1.42	1.36	1.64	1.48	1.24	1.46
	KP	2.74	2.48	3.82	3.62	3.04	1.22	2.28	1.70	1.32	1.34	1.30	1.30	1.20	1.38
	Avg	3.75	3.69	4.23	3.63	3.73	2.99	2.96	2.55	2.48	2.39	2.25	1.87	2.26	2.22
Twist	VC	4.52	4.52	4.68	4.24	4.3	4.52	2.56	2.22	2.62	2.02	1.63	1.43	1.84	1.66
	VQ	4.90	5.00	4.86	4.52	4.32	4.98	4.74	4.24	4.02	4.45	4.55	3.31	4.12	4.25
	IF	3.86	3.72	4.22	3.99	4.14	1.78	3.00	2.38	1.76	1.61	1.59	1.82	1.12	1.46
	KP	3.80	3.94	3.80	3.77	3.88	1.78	2.80	2.28	1.62	1.65	1.61	1.71	1.10	1.46
	Avg	4.27	4.30	4.39	4.13	4.16	3.27	3.28	2.78	2.51	2.43	2.35	2.07	2.05	2.21
Stretch	VC	4.36	4.34	4.26	3.80	3.98	4.20	3.22	2.38	2.36	2.07	1.90	1.54	2.10	1.51
	VQ	5.00	4.98	4.90	4.72	4.54	4.72	4.52	3.98	4.14	4.54	4.59	3.10	4.10	3.79
	IF	4.44	3.84	4.06	3.96	3.96	2.00	3.44	2.60	2.18	2.33	2.55	2.06	1.08	1.04
	KP	4.52	3.76	4.10	4.18	4.10	2.12	3.48	2.50	2.10	2.48	2.71	2.04	1.08	1.21
	Avg	4.58	4.23	4.33	4.19	4.15	3.26	3.67	2.87	2.70	2.85	2.94	2.19	2.09	1.89
Causal	VC	4.26	4.46	4.56	4.18	4.22	3.96	3.20	2.68	2.20	3.54	2.10	2.08	2.50	1.86
	VQ	4.88	4.84	4.98	4.76	4.84	4.76	4.16	4.66	3.84	4.88	4.71	3.66	4.28	4.12
	IF	4.74	4.58	4.78	4.35	4.04	2.50	3.50	2.54	2.18	1.56	1.88	2.24	1.74	1.80
	KP	4.68	4.56	4.84	4.13	3.96	2.70	3.38	2.52	2.10	1.56	1.90	2.20	1.72	1.96
	Avg	4.64	4.61	4.79	4.36	4.27	3.48	3.56	3.10	2.58	2.89	2.65	2.55	2.56	2.44
Overall	Avg	4.36	4.22	**4.45**	4.12	4.07	3.21	3.35	2.76	2.52	2.51	2.46	2.09	2.44	2.14

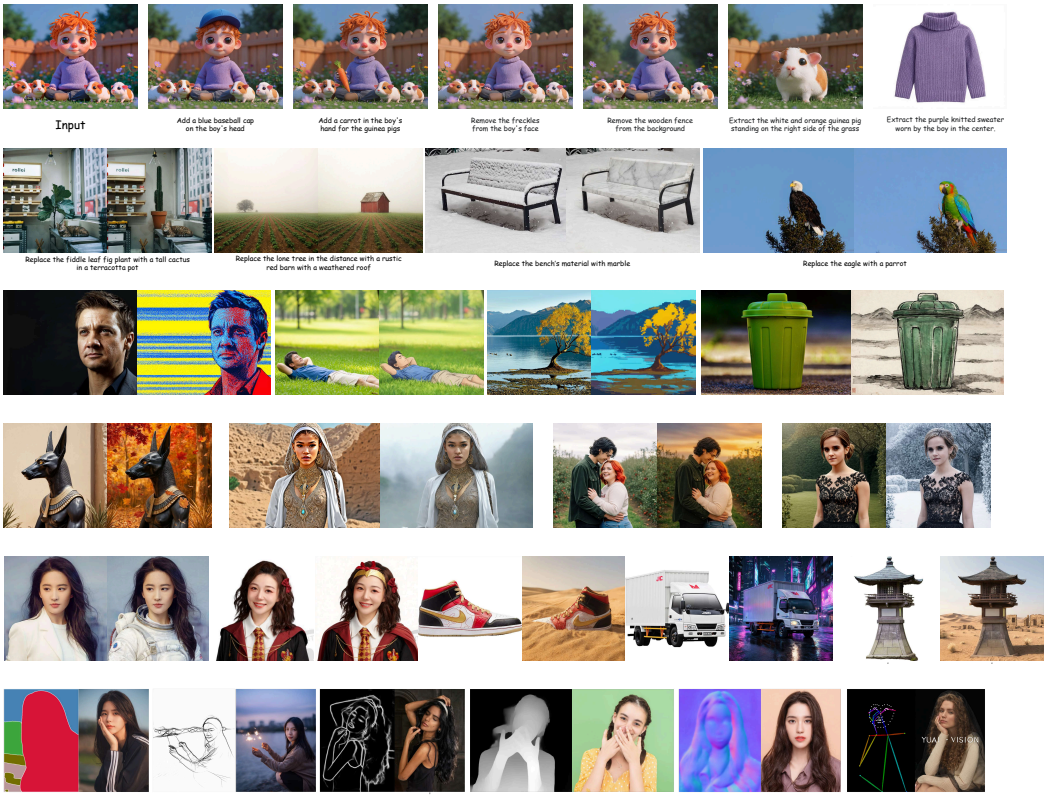


Figure 17: **More results on single image editing.** From top to bottom: (1) *Add/Remove/Extract*: adding objects, removing elements, and extracting subjects; (2) *Replace*: substituting objects or materials while preserving context; (3) *Style Transfer*: transforming image appearance into different artistic styles; (4) *Background*: modifying or replacing image backgrounds; (5) *Subject-Driven Generation*: generating new images conditioned on reference subjects; (6) *Controllable Generation*: generating images with spatial or structural control signals.

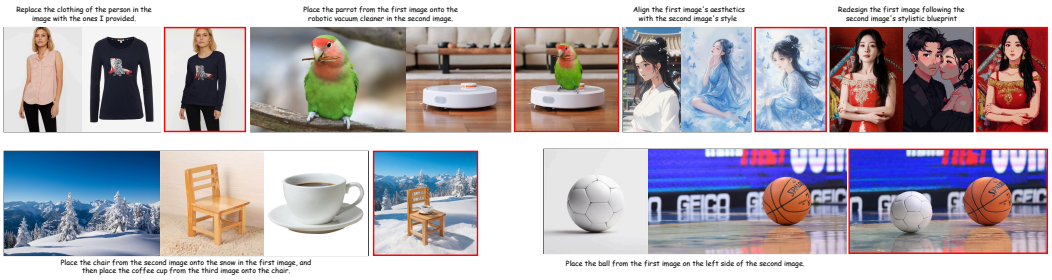


Figure 18: **More results on multi-element composition.** Our model can compose multiple reference subjects into a coherent scene while preserving the identity of each element.

J Chain-of-Thought Enhanced Generation

To further improve the quality and controllability of image editing and generation, we incorporate chain-of-thought (CoT) reasoning into the training pipeline of UNICOM. Specifically, during training, we enable CoT for a proportion of samples by prepending a `<think>` token before the output image. This encourages the model to first produce a structured reasoning trace that plans the visual transformation, and then generate the target image conditioned on both the original input and the reasoning content.

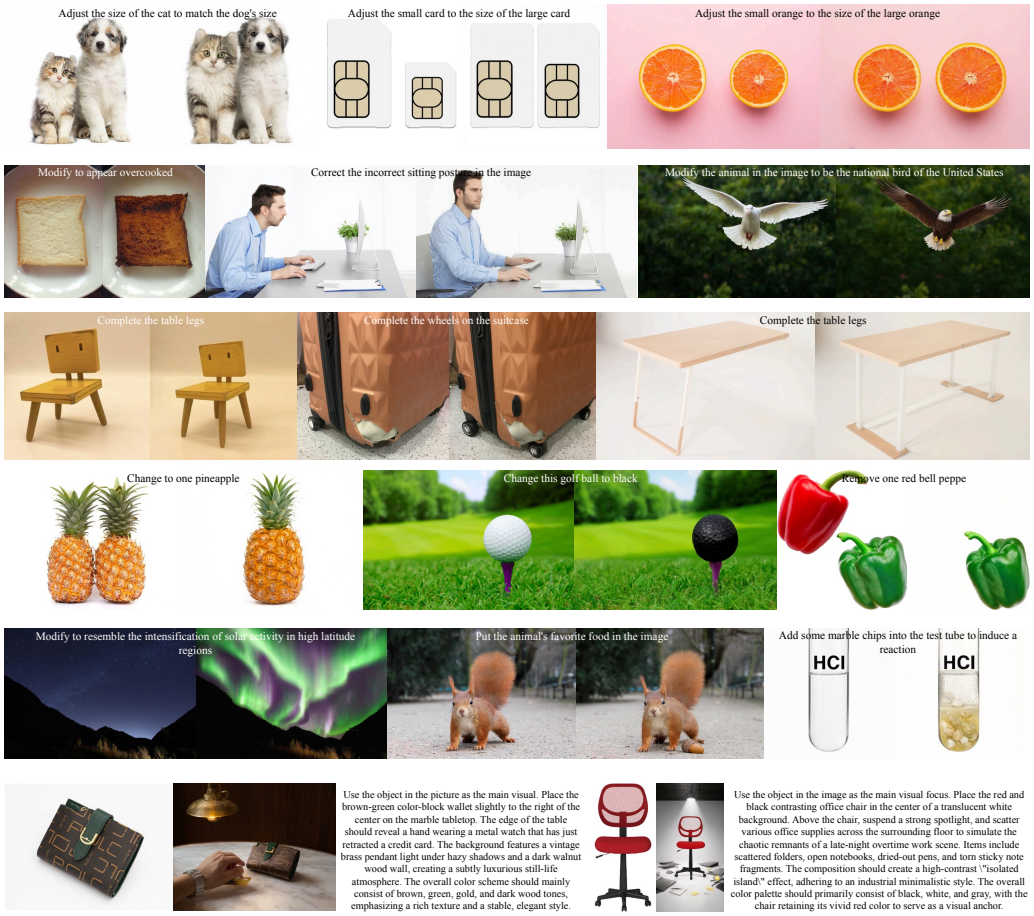


Figure 19: **More results on intelligent image editing.** These examples require complex world knowledge such as biological transformations, geographical understanding, and cultural context to perform accurate editing.

Under identical sampling settings, enabling CoT improves GenEval from 0.85 to 0.87, ImgEdit from 4.16 to 4.22, and KRIS-Bench from 62.30 to 70.11. The largest improvement, 7.81 points on KRIS-Bench, indicates that the benefit is most pronounced for knowledge-intensive editing in this evaluation.

Figure 20 presents representative examples where CoT is activated during inference. For each case, we display the generated image alongside the corresponding thinking process. As shown, the model explicitly reasons about the required modifications (e.g., identifying which attributes to change, understanding spatial relationships, or decomposing complex instructions into sub-steps) before producing the output. This explicit reasoning step enables more accurate and semantically faithful generation, particularly for tasks that involve multi-step transformations or require fine-grained control over the output.

K Related Work

The visual tokenizer [44, 39, 23] stands as a cornerstone in the realization of unified multimodal models. Its paramount function resides in transmuting visual imagery into visual tokens amenable to processing by generative models, thereby facilitating the concurrent support of both image understanding and generation tasks. Traditional visual generation has long been predicated on pixel-based encoders rooted in Variational Autoencoders (VAEs [49]), which support image generation by encoding visual data into a low-dimensional latent space. Nevertheless, such pixel-centric encoders are

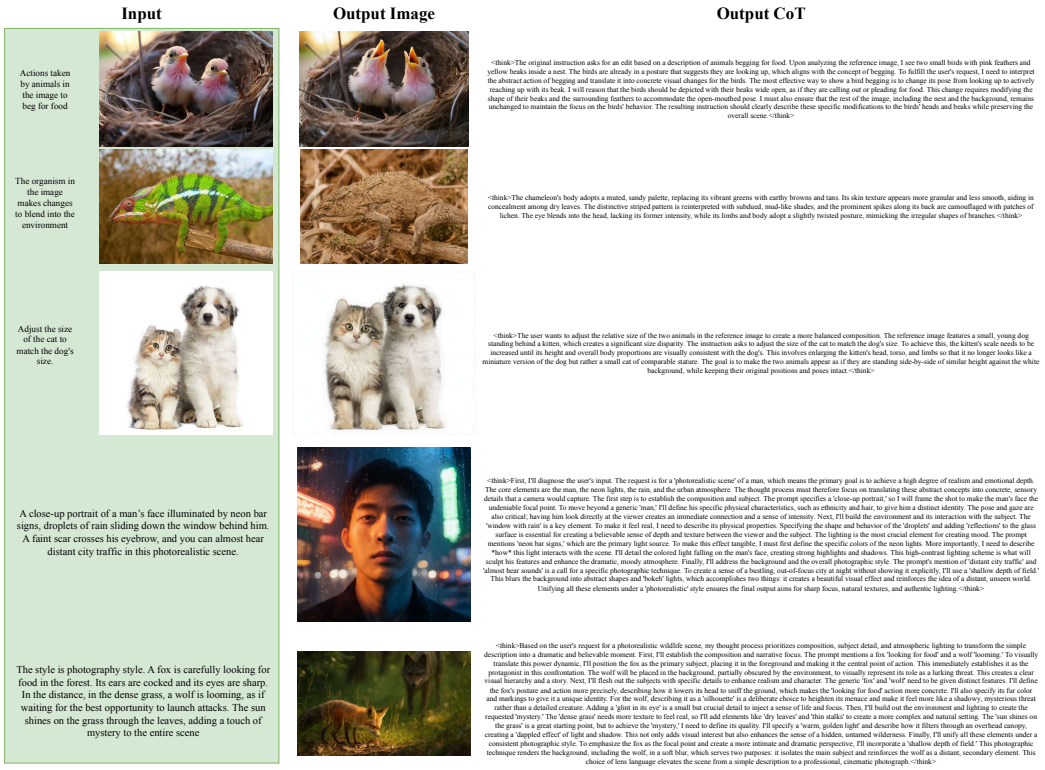


Figure 20: **Chain-of-thought enhanced generation and editing.** For each example, we show the generated image along with the corresponding thinking process enclosed in `<think>` tokens. The model first reasons about the required visual transformation, then generates the output image accordingly.

often deficient in high-level semantic features and fail to adequately harness the priors inherent in visual understanding, a limitation that constrains their performance in generative endeavors. To remediate this shortcoming, several lines of research [10, 60, 47, 71, 13] have advanced Hybrid Encoding methodologies, which integrate pixel encoding (VAE) with Transformer-based Vision Transformer (ViT) models in a bid to unify image generation and understanding. Yet, this approach remains flawed: generation and understanding tasks still operate within disparate representational spaces, engendering a disjunction between the two and compromising overall efficiency.

More sophisticated paradigms pivot toward ViT-based visual tokenizers [20, 39, 23]. For instance, Vector Quantization (VQ)-based methods [16] generate discrete visual tokens by quantizing continuous ViT features. While such approaches can effectively convey image information to generative models, the quantization process inevitably incurs information loss, and discrete token systems struggle to articulate the nuanced complexities of intricate visual features. By way of contrast, recent work has gravitated toward the employment of continuous ViT features for visual representation, exemplified by frameworks such as MingTok [25] and VUGEN [9]. These methodologies eschew the discretization process; by adopting continuous visual representations, they succeed in unifying understanding and generation tasks within a shared representational space. They surmount the twin pitfalls of information loss and task disjunction that plagued traditional methods, thereby elevating both the quality and diversity of generated images.

L Limitation

While UNICOM represents a significant step forward in unifying visual understanding and generation, there are still several limitations that need to be addressed. One primary limitation is the complexity involved in training high-dimensional continuous representations. Despite the success of our compression techniques, there is still potential for information loss during the compression

process, particularly when dealing with very fine-grained details in complex images. While we have demonstrated the ability of UNICOM to preserve these details, the trade-off between compression and fidelity remains a challenging aspect of our approach.

Another limitation arises from the inherent challenges of training generative models, UNICOM still requires substantial computational resources, especially for large-scale datasets and high-resolution image tasks. The energy consumption associated with training such large models may also limit the accessibility and sustainability of this approach in resource-constrained environments. Future work will need to address these scalability concerns, possibly through model distillation or optimization techniques.

M Broader Impacts

UNICOM advances unified visual understanding and generation, which can benefit creative content production, knowledge-intensive editing, accessibility tools, and multimodal assistants. The compressed semantic representation may also improve training efficiency, reducing the cost of building unified generative systems.

At the same time, stronger image generation and editing models can be misused to create misleading or manipulated visual content. Responsible deployment should therefore include provenance mechanisms, content moderation, watermarking or detection tools when appropriate, and clear usage policies for downstream applications.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the core claims: compressed continuous semantic representations, an attention-based semantic compressor, and a Transfusion-style predictor for unified understanding and generation. These claims are supported by reconstruction, generation, editing, and ablation results in Sections 4.2 and 4.3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The appendix includes a dedicated Limitation section discussing compression fidelity, training complexity, compute requirements, and scalability.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper is empirical and does not present formal theoretical results or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No]

Justification: We describe the architecture, training stages, evaluation protocols, and ablation setup in Sections 4.1 and 4.3, and Appendix A. Exact reproduction is limited because some training data and full-scale training infrastructure are not released with the anonymous submission.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Upon acceptance of this paper, we will release the code, model checkpoints, and data preparation instructions needed to reproduce the main results. The anonymous submission provides methodological and experimental details in the paper and appendix.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the model components, training stages, default noise shift, evaluation benchmarks, and ablation settings in Sections 4.1 and 4.3; additional hyperparameters and data mixtures are given in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars or statistical significance tests. The main experiments are large-scale generative evaluations where repeated full training runs are computationally expensive.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix A reports the training stages, batch size, optimizer, data mixtures, hardware, and time cost in Table 15. The main paper also reports compute savings for the key compression ablation in Section 4.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work studies unified multimodal modeling and does not involve human-subject experiments, private user data release, or deceptive deployment. We have reviewed the NeurIPS Code of Ethics and aim to comply with it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The appendix includes a Broader Impacts section discussing potential benefits for creative and accessibility applications as well as misuse risks such as misleading or manipulated visual content.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The anonymous submission does not release model checkpoints, generated datasets, or a public deployment. Therefore, release safeguards are not applicable at submission time.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Existing models, datasets, and benchmarks are cited where used, including SigLIP2, FLUX, Qwen, ImageNet, and evaluation benchmarks. We do not redistribute these assets in the submission.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release new datasets, code packages, or model checkpoints as part of the anonymous submission.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing experiments or human-subject research, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large language models such as Qwen are used as the foundational backbone for the multimodal architecture, which is a standard and expected paradigm in this research area rather than a non-standard methodology requiring a separate LLM-usage declaration.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.