HIUNI: HARMONIZING UNDERSTANDING AND GENERATION IN SHARED VISUAL SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models (UMMs) aim to support visual understanding and generation within a single framework. However, prevailing approaches often separate their visual representations or backbone computation to accommodate the different requirements of the two tasks, which limits the degree of unification. In this paper, we present HIUNI, which revisits this design choice and explores whether a pretrained VLM can be directly extended into a unified model with a high degree of sharing. Specifically, HIUNI reuses the same language aligned ViT and dense Transformer backbone for both capabilities, while confining task specific computation to the output side. Yet such a design introduces two key challenges. Direct generation in pixel space is difficult to optimize, while visual compression and semantic abstraction in pretrained VLMs may discard fine spatial details required for synthesis. To address these challenges, HIUNI introduces a Pixel Diffusion Head tailored to direct pixel-space optimization and a lightweight HiGate pathway that provides fine-grained spatial details near the output stage. Extensive experiments show that HIUNI achieves a similar average understanding score to the pretrained VLM while attaining competitive performance on image generation and editing. Controlled comparisons further demonstrate that a single dense backbone can effectively support both capabilities without task-specific backbone computation, highlighting a simpler path toward unified multimodal modeling.

1 INTRODUCTION

Image understanding translates visual content into language, whereas visual generation synthesizes images from textual descriptions. An ideal unified multimodal model (UMM) is expected to support both capabilities within a single framework and facilitate cross-task interaction between them (Kang et al., 2026; Wu et al., 2026). However, most existing UMMs still maintain two separate visual representation spaces with dedicated encoders. The understanding branch extracts high-level semantics via a language-aligned Vision Transformer (ViT), while the generation branch learns fine-grained spatial details with a Variational Autoencoder (VAE). As highlighted by Janus (Wu et al., 2025b), forcing two divergent encoding objectives into one representation triggers severe task interference. Bagel (Deng et al., 2025) alleviates this conflict through a Mixture-of-Transformers (MoT) architecture that separates task-specific computation within the backbone, establishing a modular paradigm adopted by recent studies (Agarwal et al., 2026; Diao et al., 2026; Yan et al., 2026).

Although this separation preserves individual performance, it limits direct interaction through shared visual representations. Recent studies further suggest that closer sharing of visual representations can promote stronger interaction between the two capabilities with an appropriate backbone design (Zheng et al., 2025; Su et al., 2026). These observations motivate us to revisit whether such architectural separation is indeed necessary: *Can a shared visual space internally harmonize understanding and generation without relying on task-specific backbone computation?*

Recent advances provide a timely opportunity to answer this question. Specifically, a straightforward path to eliminating the dual-space discrepancy is to represent both tasks with a shared visual pathway (Liu et al., 2026b; Zhang et al., 2026a). Conventional generation models heavily rely on VAE latents, yet these features are optimized primarily for reconstruction and are less aligned with the high-level semantics required for understanding (Wu et al., 2025e; Ma et al., 2026). Conversely, modern pixel-space diffusion models demonstrate that high-quality generation can be achieved with-

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

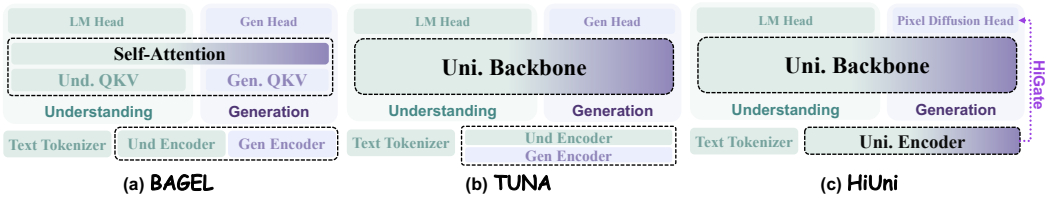


Figure 1: Comparison of representative UMM designs. (a) Bagel utilizes separate visual encoders and task-specific MoT modules for understanding and generation. (b) TUNA constructs a shared visual representation by cascading VAE and ViT. (c) HiUni reuses the pretrained VLM visual pathway, while *HiGate* supplies fine-grained pixel details only to the generation head.

out VAE latents (Li & He, 2026; Yu et al., 2026), opening a new path towards shared visual representations. Going one step further, some encoder-free architectures process raw pixels for both understanding and generation (Liu et al., 2026a; Diao et al., 2026), but the absence of pretrained visual semantic priors makes optimization substantially less stable and less data efficient. An alternative choice is to reuse a language-aligned ViT as the shared visual encoder, given its strong semantic representations and direct compatibility with standard VLM architectures. Nevertheless, adapting such an encoder to image generation introduces a spatial bottleneck. Standard VLMs typically employ a patch merger after ViT to compress visual tokens, doubling effective patch size from $14/16$ to $28/32$ (Bai et al., 2025; Guo et al., 2025; Wang et al., 2025). Such aggressive spatial pooling, together with the semantic abstraction of deep ViT layers, improves understanding efficiency, but weakens fine-grained information needed for high-fidelity synthesis.

Following this direction, we introduce HiUni, a UMM initialized from a pretrained VLM that extends image understanding to generation through a shared semantic visual pathway. Both understanding and generation operate on the same visual tokens and share the same backbone, without introducing additional encoders. Task-specific computation is confined to the output side. For image generation, we observed optimization instability when applying a standard DiT to pixel-space diffusion in our early experiments. We therefore design a dedicated *Pixel Diffusion Head* with a simplified modulation for pixel-space generation. To compensate for spatial information weakened by patch merging and deep semantic abstraction, we incorporate a lightweight gating module (*HiGate*), inspired by Cheers (Zhang et al., 2026b), which directly routes uncompressed pixel-level features to the pixel diffusion head. This residual pathway supplies local spatial cues for generation while leaving semantic modeling in the pretrained visual encoder and shared backbone.

We validate HiUni on image understanding, image generation, and image editing. Despite incorporating image generation, HiUni maintains strong visual understanding across a broad set of benchmarks. Meanwhile, HiUni achieves strong generation and editing performance, including the strongest overall Chinese score on Qwen-Image-Bench among the unified models compared, together with competitive image editing results. Under matched training settings, task-routed MoT improves generation over a same-scale dense backbone. Separately, a larger shared dense backbone with a similar total parameter count achieves substantially stronger understanding while retaining competitive generation, indicating that scaling a shared dense backbone within a shared visual space offers a simple and promising path toward unified multimodal modeling.

Our contributions are summarized as follows:

- We propose HiUni, which offers a new perspective on building UMMs from pretrained VLMs. Understanding, generation, and editing share the same visual encoder and dense backbone, while computation specific to each task is limited to the output side.
- We develop a *Pixel Diffusion Head* tailored to direct pixel-space generation, together with a lightweight residual pathway that complements the shared semantic representation with fine-grained spatial details.
- Extensive experiments show that HiUni maintains strong understanding while achieving competitive generation and editing performance. Controlled comparisons further show that a shared dense backbone can effectively support both understanding and generation without task-specific backbone computation.

2 Related Work

Unified Multimodal Models. UMMs are intended to support visual understanding and generation within a single model. Existing UMMs mainly follow two paradigms according to how images are represented and generated. Autoregressive models tokenize images into discrete visual tokens and model text and vision under a unified autoregressive objective (Team, 2024; Cui et al., 2025; Wu et al., 2025e; Team et al., 2026; Wang et al., 2026b). This formulation provides a simple and scalable interface across modalities, but discrete tokenization may lose subtle visual details required for high fidelity image synthesis. Another line of work follows the Transfusion paradigm (Zhou et al., 2025), which retains the autoregressive prediction of language while applying diffusion or flow matching to continuous visual features for image generation (Deng et al., 2025; Liu et al., 2026b; Yan et al., 2026). More recently, pixel-space diffusion further demonstrates that strong image synthesis can be achieved without VAE latents (Li & He, 2026; Yu et al., 2026), opening a more direct path toward shared visual representations. Together, these advances shift an important architectural question toward how visual representations and backbone computation should be shared between understanding and generation.

Visual Unification. Many existing models maintain separate visual spaces due to the different requirements of the two tasks, where a language-aligned ViT provides semantic features for understanding, while a reconstruction-oriented VAE preserves fine-grained details for generation (Wu et al., 2025b; Deng et al., 2025). To reduce this separation, recent studies explore more unified visual pathways. TUNA cascades a VAE encoder with a representation encoder to construct a shared visual representation, while Tuna-R further removes the VAE for pixel space modeling (Liu et al., 2026b;a). HYDRA adopts a progressive ViT, and UniSpace introduces reconstruction aware patch representations into a pretrained ViT (Zhang et al., 2026a; Yan et al., 2026). These designs reduce visual separation, but understanding and generation may still rely on different encoder stages or task-specific backbone computation. For example, UniSpace still adopts task-specific Transformer experts, while encoder-free models such as SenseNova remove both ViT and VAE but retain an MoT backbone (Yan et al., 2026; Diao et al., 2026). In contrast, HiUni directly reuses the language-aligned ViT and patch merger of a pretrained VLM, shares the resulting visual tokens and a single dense backbone across understanding and generation.

3 Method

Image understanding and generation differ substantially in their input and output forms, but both rely on semantic alignment between visual content and language. HiUni therefore separates shared semantic modeling from task-specific prediction. Specifically, we seek to let the pretrained visual encoder and multimodal backbone provide a common semantic pathway for both capabilities, while introducing generation-specific processing only when detailed visual signals need to be reconstructed. This design allows understanding and generation to share the main visual and backbone computation without requiring separate semantic representations.

Following this principle, HiUni is built upon a pretrained VLM and extends it with image generation capability while reusing its pretrained visual encoder and backbone architecture. Both understanding and generation use the same language-aligned ViT, the same visual tokens after patch merging, and the same backbone. Textual outputs are predicted by the original language modeling head, while image synthesis is handled by a new pixel diffusion head. No separate semantic visual encoder or task-specific backbone is involved. To compensate for the spatial information weakened by patch merging and deep semantic abstraction, HiUni further adopts a lightweight gating module, termed *HiGate*, which routes raw features from the noisy image to the diffusion head. The residual branch supplies local spatial cues to complement the shared semantic representation without modifying the main semantic computation. Figure 2 illustrates the overall architecture.

3.1 Unified Multimodal Architecture

Shared Visual Encoding. HiUni adopts a language aligned ViT from the pretrained VLM as the only visual encoder for both understanding and generation. Let $\mathcal{E}$ denote the ViT and $\mathcal{M}$ denote the patch merger, which compresses the ViT outputs into visual tokens. For an input image $\mathbf{x}$, the visual tokens are obtained as $\mathbf{Z} = \mathcal{M}(\mathcal{E}(\mathbf{x}))$. The same pathway is used for clean images in understanding

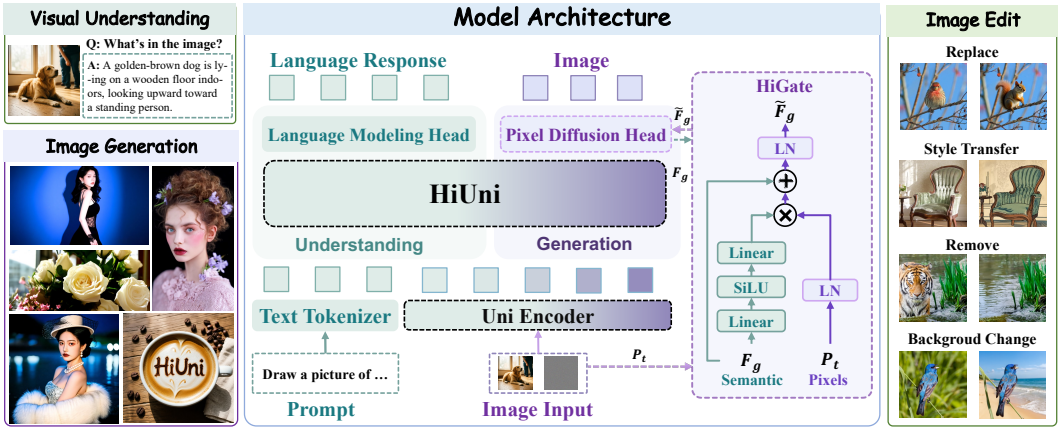


Figure 2: Overview of HIUNI. Starting from a pretrained VLM, it shares the same visual encoder and dense backbone across understanding, generation, and editing. Image synthesis is handled by a *Pixel Diffusion Head*, while *HiGate* provides a lightweight residual pathway that supplies local spatial cues weakened by visual compression.

and visual conditioning, as well as noisy target images in generation. Consequently, both tasks are grounded in the same visual token space without requiring separate or task-specific encoders.

Shared Dense Backbone. The shared visual tokens are further processed by the same dense backbone for both understanding and generation. Text embeddings and visual tokens are organized into a multimodal sequence $\mathbf{S}$ and processed by a single dense Transformer $\mathcal{B}$, yielding $\mathbf{H} = \mathcal{B}(\mathbf{S})$. The attention mask allows bidirectional interaction within each image span, while text tokens retain causal attention for autoregressive prediction. After the shared backbone, task specific prediction is handled on the output side. For image understanding, hidden states corresponding to textual outputs are projected to the vocabulary space by the language modeling head. For image generation, hidden states at the target visual positions are passed to the pixel diffusion head for image prediction.

3.2 PIXEL-SPACE IMAGE GENERATION

Pixel Diffusion Head. Diffusion directly in pixel space poses additional optimization challenges compared with conventional latent diffusion, as the model must predict high dimensional pixel signals while maintaining semantic consistency and fine spatial structure (Jiang et al., 2026; Li & He, 2026). Consistent with these observations, we observed frequent loss spikes when using a standard DiT (Peebles & Xie, 2023) with repeated AdaLN modulation in our early experiments. In subsequent development runs, a simplified design without block wise AdaLN exhibited smoother training trajectories. Motivated by this observation, we remove AdaLN from the DiT blocks, reducing the diffusion head to a standard Transformer architecture.

Removing AdaLN does not mean removing timestep conditioning from the model. Inspired by Transfusion (Zhou et al., 2025), we represent the timestep as an explicit time token. For timestep t, a sinusoidal encoding $\phi(t)$ is mapped by a two layer MLP to a time embedding $\mathbf{e}_t \in \mathbb{R}^{D_b}$, where D_b denotes the hidden dimension of the backbone. The resulting vector replaces the embedding of the `<time>` token and is processed together with the multimodal sequence by the shared backbone. The backbone hidden states for the target visual positions, together with the time token representation, are projected to the diffusion hidden dimension D and processed by N Transformer blocks.

At the output of the diffusion head, we retain a single timestep modulation before pixel prediction. A separate two layer MLP maps the same sinusoidal encoding $\phi(t)$ to a diffusion time embedding $\mathbf{c}_t \in \mathbb{R}^D$. This embedding is further projected to channel wise scale and shift parameters as $[\boldsymbol{\gamma}_t, \boldsymbol{\beta}_t] = \mathbf{W}_{\text{mod}}\,\text{SiLU}(\mathbf{c}_t) + \mathbf{b}_{\text{mod}}$. Let $\mathbf{F}_{\text{out}} \in \mathbb{R}^{N_g \times D}$ denote the target visual features at the output of the diffusion head. The final modulation is given by $\bar{\mathbf{F}}_{\text{out}} = (1+\boldsymbol{\gamma}_t) \odot \text{LN}(\mathbf{F}_{\text{out}}) + \boldsymbol{\beta}_t$. After modulation, a linear projection predicts the corresponding pixel patches. With a ViT patch size of 16 and a spatial merge factor of 2, each generated token predicts a 32×32 RGB region, corresponding to a flattened

patch dimension of $N_p = 3 \times 32 \times 32$. Finally, the predicted patches are rearranged into image space to obtain $\hat{\mathbf{x}}_0$. In this way, the head removes repeated block-wise AdaLN while preserving timestep conditioning through the shared backbone and final modulation layer.

HiGate. Although the unified visual tokens carry strong semantics, the patch merger reduces their spatial resolution, while deeper ViT layers increasingly favor abstraction over local texture. Such compression is well suited to understanding but may discard cues important for pixel-level reconstruction. To complement the semantic pathway, HiUni introduces a lightweight *hi-gate* module, inspired by Cheers (Zhang et al., 2026b), which directly injects low-level pixel information into the diffusion head without introducing an additional visual encoder.

Specifically, the noisy image $\mathbf{x}_t \in \mathbb{R}^{3\times H\times W}$ branches into two pathways before ViT patch embedding. The main pathway encodes $\mathbf{x}_t$ through the ViT, patch merger, and shared backbone to obtain semantic features, while a lightweight residual branch directly preserves its local pixel content. In the residual branch, each 2×2 group of 16×16 RGB patches is concatenated into $\mathbf{R}_t \in \mathbb{R}^{N_g\times N_p}$, following the same spatial layout as the target visual tokens. A linear projection maps these raw pixel features to the diffusion hidden dimension D, producing $\mathbf{P}_t \in \mathbb{R}^{N_g\times D}$. The aligned pixel features $\mathbf{P}_t$ are then fused with the target features $\mathbf{F}_g \in \mathbb{R}^{N_g\times D}$ through *HiGate*. A lightweight MLP predicts a token wise gate $\mathbf{g}_t \in \mathbb{R}^{N_g\times 1}$, and the residual fusion is formulated as $\widetilde{\mathbf{F}}_g = \mathrm{LN}(\mathbf{F}_g + \mathbf{g}_t \odot \mathrm{LN}(\mathbf{P}_t))$. Only the target visual positions are updated by this residual injection, and the resulting features continue through the *Pixel Diffusion Head*. In this manner, the residual pathway supplies fine-grained spatial cues for image synthesis, while high-level semantic modeling remains in the unified ViT and shared backbone.

3.3 Task Formulation and Joint Training

Unified Task Formulation. HiUni supports text-to-image generation and image editing through the same generation pathway, with the two tasks differing only in their conditioning context. For image generation, the condition is a textual prompt $\mathbf{p}$. For image editing, the condition consists of an editing instruction $\mathbf{p}$ and a reference image $\mathbf{x}_r$. The reference image follows the same visual encoding pathway as the target image and is processed together with the noisy target image by the shared backbone. Formally, we denote the conditioning context by $\mathbf{C}_\tau$, where $\tau \in \{\mathrm{t2i}, \mathrm{edit}\}$, $\mathbf{C}_{\mathrm{t2i}} = \mathbf{p}$, and $\mathbf{C}_{\mathrm{edit}} = (\mathbf{p}, \mathbf{x}_r)$. Given $\mathbf{C}_\tau$ and the noisy target image $\mathbf{x}_t$, both tasks follow the same generation pathway without introducing task specific branches.

Joint Training Objectives. HiUni jointly optimizes language modeling, image understanding and image generation within the same backbone. Textual outputs for language modeling and image understanding are trained with the standard next token prediction objective. Let $\mathbf{C}$ denote the input context, the language modeling objective is $\mathcal{L}_{\mathrm{txt}} = -\frac{1}{M}\sum_{i=1}^{M} \log p_\theta(y_i \mid \mathbf{C}, y_{<i})$. For image generation, we follow JiT (Li & He, 2026) and adopt x-prediction with a velocity-based objective. Given a clean target image $\mathbf{x}_0$, Gaussian noise $\mathbf{z} \sim \mathcal{N}(0, I)$, and a timestep t sampled such that $\mathrm{logit}(t) \sim \mathcal{N}(-0.8, 0.8^2)$, we define $\boldsymbol{\epsilon}_{N_g} = \sigma_{N_g}\mathbf{z}$, where σ_{N_g} is a resolution dependent noise scale detailed in Appendix A. The noisy image is constructed as $\mathbf{x}_t = t\mathbf{x}_0 + (1-t)\boldsymbol{\epsilon}_{N_g}$. The model takes $\mathbf{x}_t$, timestep t, and task condition $\mathbf{C}_\tau$ as input and directly predicts the clean image as $\hat{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t, t, \mathbf{C}_\tau)$. The target velocity is $\mathbf{v} = \mathbf{x}_0 - \boldsymbol{\epsilon}_{N_g}$. For numerical stability near $t = 1$, the predicted velocity is computed as $\hat{\mathbf{v}} = (\hat{\mathbf{x}}_0 - \mathbf{x}_t)/\max(1-t, 0.05)$. The generation objective is then defined as $\mathcal{L}_{\mathrm{gen}} = \mathbb{E}[\|\hat{\mathbf{v}} - \mathbf{v}\|_2^2]$. The two objectives are jointly optimized as $\mathcal{L} = \mathcal{L}_{\mathrm{txt}} + \lambda_{\mathrm{gen}}\mathcal{L}_{\mathrm{gen}}$, where λ_{gen} balances the two objectives.

Training stages. HiUni is trained in three stages consisting of generation warmup, joint pretraining, and supervised fine tuning. During warmup, we keep the pretrained backbone and visual encoder frozen, and optimize only the newly initialized *Pixel Diffusion Head* and *HiGate*. This stage first establishes the mapping from pretrained VLM representations to pixel outputs without perturbing the pretrained components. Next, we unfreeze the entire model and jointly train text modeling, image understanding, and image generation. The image resolution is progressively increased during this stage, allowing generation to gradually adapt from low to high resolutions. Finally, supervised fine tuning introduces image editing while retaining the existing tasks.

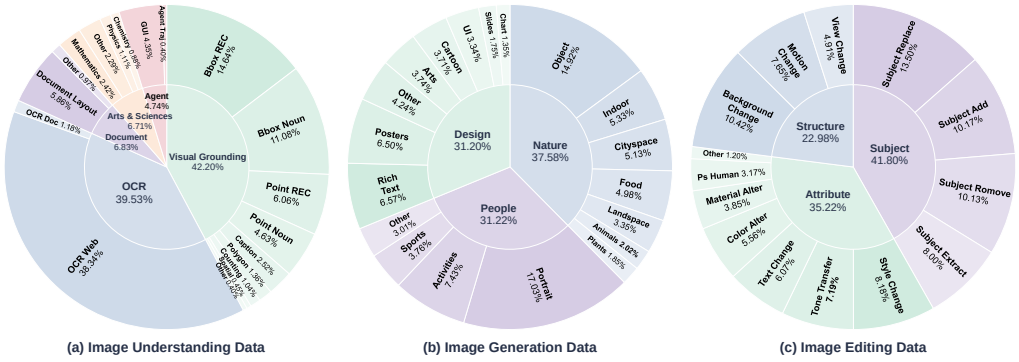


Figure 3: Data distribution of HiUni's training corpus for image understanding, generation, and editing. Inner rings show the major categories, while outer rings provide their finer-grained breakdown.

4 EXPERIMENTS

We evaluate HIUNI on image understanding, image generation, and image editing, and further examine architectural design through controlled comparisons. Additional training and evaluation details, analyses of pixel diffusion optimization stability and cross task interaction, an extension to chain of thought image generation, and qualitative results are provided in Appendices A-F.

4.1 EXPERIMENTAL SETUP

Implementation Details. We initialize HIUNI from Qwen3-VL-4B-Instruct (Bai et al., 2025), reusing its language model backbone as the shared dense backbone and its language aligned ViT as the shared visual encoder. The multimodal training corpus covers image understanding, image generation, and image editing, with their category distributions summarized in Figure 3. Following the three-stage procedure in Section 3.3, we first warm up the *Pixel Diffusion Head* and *HiGate* for 30K steps on image generation data while freezing all pretrained components. We then unfreeze the full model and perform 310K steps of joint pretraining on pure text modeling, image understanding, and image generation, with the generation resolution progressively increased from 256×256 to 2048×2048. Finally, we conduct 22K steps of supervised fine tuning, where image editing is added while the existing tasks are retained. The learning rate is 1×10^{-4} during generation warmup and early joint pretraining, and is reduced to 2×10^{-5} in later pretraining, while supervised fine tuning uses a peak learning rate of 2×10^{-5}. The complete phase-wise recipe, including training steps, batch sizes, learning rate schedules, image resolutions, loss weights, and task sampling ratios, is provided in Appendix A and Table 6.

Baselines. We compare HIUNI with a broad set of united and task-specific models. Representative UMM baselines include Harmon (Wu et al., 2025d), Cheers (Zhang et al., 2026b), InternVL-U (Tian et al., 2026), Janus-Pro (Chen et al., 2025), Show-o2 (Xie et al., 2026), TUNA (Liu et al., 2026b), Bagel (Deng et al., 2025), SenseNova-U1 (Diao et al., 2026), UniSpace (Yan et al., 2026) and among others. For image understanding, we also report Qwen3-VL-4B-Instruct (Bai et al., 2025), from which HIUNI is initialized, as a reference for the image understanding capability of the pretrained VLM. For image generation, we further compare against strong baselines such as GLM-Image (GLM-Image Team, 2026), Qwen-Image (Wu et al., 2025a), and HiDream-O1 (Cai et al., 2026). For image editing, we include recent specialized models Qwen-Image-Edit-2511 (Qwen Team, 2025), LongCat-Image-Edit (Meituan LongCat Team, 2026), and Qwen-Image-2.1 (Qwen Team, 2026). As many earlier benchmarks are nearly saturated, we focus on newer and more discriminative benchmarks here, so the compared models may vary across tables.

4.2 MAIN RESULTS

Image Understanding. We evaluate HIUNI on eight representative benchmarks, including MathVista (Lu et al., 2024), MMBench (Liu et al., 2024a), RealWorldQA (AI, 2024), MMStar (Chen

Table 1: Comparison with representative models on image understanding. HiUni is initialized from Qwen3-VL_{INS}-4B. The asterisk * indicates official results. Model size is reported as understanding (backbone) parameters + generation specific parameters. Hereinafter the same.

Models	Size	MathVista$_{Mini}$	MMBench$_{EN}$	RealWorldQA	MMStar	HallusionBench	AI2D	OCRBench	OmniDocBench ↓
Understanding-Only Models									
Qwen3-$VL^\star_{INS}$	4B	73.7	85.1	70.9	69.8	57.6	84.1	88.1	0.24
Qwen3-$VL^\star_{THK}$	4B	79.5	84.6	73.2	73.2	64.1	84.9	80.8	0.23
Unified Multimodal Models									
Harmon*	1.5B	–	65.5	49.8	35.3	–	–	11.2	–
Cheers*	1.5B	50.5	70.4	60.9	50.9	–	–	58.4	–
InternVL-U	2B + 1.7B	53.9	71.7	56.9	53.5	32.4	76.8	83.9*	0.83
Janus-Pro	7B	44.7	79.2*	58.0	48.3	36.4	71.3	59.0	0.81
Show-o2	7B	51.8	79.3*	64.7*	56.6	36.2	78.6*	32.4	0.84
TUNA*	7B	–	–	66.1	61.2	–	79.3	74.3	–
Bagel	7B + 7B	73.1*	85.0*	74.4	67.6	51.8	89.2	80.8	0.43
SenseNova-U1	8B + 8B	84.2*	90.3*	73.6*	78.3*	67.8*	91.7*	82.1*	0.28
HiUni	4B + 1.3B	79.0	80.7	72.2	70.5	66.0	86.0	74.5	0.24

Table 2: Comparison with representative models on image generation. More results of OneIG are provided in Table 8. **Bold** denotes the best result among UMMs, while underline denotes the second-best result. Hereinafter the same.

Models	Size	Qwen-Image-Bench$_{ZH}$						Qwen-Image-Bench$_{EN}$						OneIG		LongText	
		Qual.	Aesth.	Align.	Real.	Creat.	Overall	Qual.	Aesth.	Align.	Real.	Creat.	Overall	ZH	EN	ZH	EN
Generation-only Models																	
GLM-Image	9B + 7B	49.26	50.64	47.90	44.69	45.23	48.19	49.86	49.98	47.49	44.25	44.67	47.86	0.511*	0.528*	0.979*	0.952*
Qwen-Image*	20B	48.44	52.25	50.72	43.16	47.30	49.23	48.45	51.18	50.04	43.45	45.37	48.48	0.548	0.539	0.946	0.943
HiDream-O1	8B	48.61	49.75	49.53	43.97	40.96	47.58	48.84	49.69	49.53	43.80	41.05	47.59	0.518	0.521	0.978*	0.979*
Unified Multimodal Models																	
InternVL-U	2B + 1.7B	42.34	43.38	40.30	38.13	34.36	40.54	43.92	43.03	41.09	38.87	32.35	40.82	0.500*	0.500*	0.860*	0.738*
Bagel	7B + 7B	46.11	45.88	40.27	39.74	32.69	42.15	45.90	47.20	43.48	40.30	34.83	43.45	0.370	0.361	0.310	0.373
SenseNova-U1*	8B + 8B	45.79	47.07	47.78	42.53	43.88	45.99	47.30	50.30	50.58	43.13	46.15	**48.28**	0.535	0.549	0.962	**0.979**
UniSpace	8B + 8B	47.22	45.02	48.54	46.19	40.49	46.05	47.44	46.54	48.94	47.54	43.84	47.05	**0.533***	**0.561***	–	–
HiUni	4B + 1.3B	46.73	50.00	50.45	43.09	45.62	**47.98**	47.63	50.52	48.78	43.70	44.15	47.70	0.524	0.538	**0.963**	0.954

et al., 2024), HallusionBench (Guan et al., 2024), AI2D (Kembhavi et al., 2016), OCRBench (Liu et al., 2024b), and OmniDocBench (Ouyang et al., 2025). As shown in Table 1, HiUni achieves strong performance across the evaluated understanding benchmarks. Across the seven higher is better benchmarks, HiUni achieves an average score of 75.56, nearly identical to the 75.61 reported for Qwen3-VL-4B-Instruct, while maintaining the same OmniDocBench score of 0.24. Performance changes are benchmark dependent, with improvements on benchmarks such as MathVista and HallusionBench and declines on others such as MMBench and OCRBench. Despite its relatively compact model scale, HiUni achieves understanding performance comparable to several substantially larger UMMs across the evaluated benchmarks. Overall, these results indicate that strong understanding performance can coexist with image generation in the proposed shared architecture.

Image Generation. We evaluate HiUni on Qwen-Image-Bench (Li et al., 2026), OneIG (Chang et al., 2026), and LongText (Geng et al., 2025). These benchmarks provide complementary evaluations that cover perceptual quality, esthetics, semantic alignment, realism, creativity, and the ability to render long text within images. From Table 2, we can observe that HiUni achieves the highest Chinese score among UMMs on QwenImageBench, while remaining competitive on the English split. For OneIG and LongText, HiUni also stays close to the strongest unified baselines, indicating strong performance on both general image generation and long text rendering. Together, these results show that HiUni achieves strong generation capability within a UMM architecture.

Image Editing For ImgEdit (Ye et al., 2026), we use GPT-5.2 as the judge model and report scores on a scale 0 to 5. For GEdit (Liu et al., 2025), we follow the VIEScore protocol (Ku et al., 2024) but replace the judge backbone with GPT-5.2, which scores semantic consistency (G_SC) and perceptual quality (G_PQ) on a scale 0 to 10. As shown in Table 3, HiUni reaches an Overall score of 3.66 on ImgEdit, ranking second among the compared unified models behind UniSpace (3.82) and slightly ahead of SenseNova-U1 (3.65), despite its relatively compact model scale. Editing-only models that are trained exclusively for this task remain ahead, reaching 4.02 and 4.01 with LongCat-Image-Edit and Qwen-Image-Edit-2511, respectively. On GEdit, HiUni obtains overall scores of 6.70 and 6.81 on Chinese and English prompts, second among the compared unified models behind SenseNova-U1 (7.05 and 6.94). These results further indicate the effectiveness of our unified architectural design across diverse and complex tasks.

Table 3: Comparison with representative models on image editing. Both benchmarks are judged by GPT-5.2. More results of GEdit are provided in Table 9.

Models	Size	ImgEdit										GEdit_{CN}			GEdit_{EN}		
		Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall	G_SC	G_PQ	G_O	G_SC	G_PQ	G_O
Editing-only Models																	
Qwen-Image-Edit-2511	20B	4.22	4.24	3.34	4.37	3.98	3.61	4.39	3.25	4.69	4.01	8.25	6.42	7.19	8.20	6.46	7.17
LongCat-Image-Edit	6B	4.15	4.24	3.39	4.47	3.95	3.57	4.28	3.39	4.70	4.02	8.13	6.44	7.15	8.05	6.44	7.10
Qwen-Image-2.1	7B	4.17	4.24	2.16	4.48	4.13	3.75	4.27	3.15	2.89	3.69	7.60	6.45	6.73	7.91	6.40	6.92
Unified Multimodal Models																	
InternVL-U	2B + 1.7B	3.94	3.71	2.42	4.08	3.34	3.45	4.03	2.28	4.07	3.48	6.02	5.81	5.74	6.05	5.80	5.74
Bagel	7B + 7B	3.49	3.36	1.17	3.69	2.82	3.06	3.80	2.59	3.82	3.09	7.37	5.71	6.31	7.31	5.68	6.27
SenseNova-U1	8B + 8B	3.82	3.96	2.44	4.20	3.11	3.59	4.47	3.09	4.20	3.65	8.19	6.20	**7.05**	8.04	6.14	**6.94**
UniSpace	8B + 8B	4.06	4.01	3.00	4.32	3.99	3.52	4.07	3.16	4.25	**3.82**	7.72	5.84	6.57	7.88	5.87	6.68
HiUni	4B + 1.3B	3.93	3.95	2.72	4.05	3.47	3.32	4.34	3.22	3.95	3.66	7.87	5.88	6.70	7.93	6.00	6.81

Table 4: Comparison between shared dense and task-routed MoT backbones under controlled settings. HiUni-Dense-2B and HiUni-MoT-2B differ only in backbone design, while HiUni-Dense-4B evaluates a larger shared backbone under a similar total parameter budget.

Backbone		Understanding							Generation
Type	Params.	$\text{MMBench}_{\text{EN}}$	RealWorldQA	MMStar	HallusionBench	AI2D	OCRBench	Avg.	OneIG Avg.
HiUni-Dense-2B	2B+0.5B	64.94	60.39	46.67	38.70	70.47	70.10	58.55	0.291
HiUni-MoT-2B	2B+2.5B	67.97	63.27	47.00	35.33	71.76	70.02	59.23	**0.345**
HiUni-Dense-4B	4B+0.5B	**72.60**	**65.49**	**53.93**	**48.96**	**77.95**	**71.30**	**65.04**	0.314

4.3 Shared versus Task-Specific Computation

Experimental Configurations. In this section, we examine whether understanding and generation require task specific backbone computation, or can be effectively supported by a shared dense backbone. We compare three configurations, HiUni-Dense-2B, HiUni-MoT-2B, and HiUni-Dense-4B. HiUni-Dense-2B and HiUni-MoT-2B are both initialized from Qwen3-VL-2B. For HiUni-MoT-2B, the pretrained backbone is duplicated to initialize separate understanding and generation branches. These two models follow the same training recipe and differ only in whether the backbone are shared or task-routed, which both undergo 30K warmup steps followed by 30K pretraining steps. Moreover, HiUni-Dense-4B is initialized from Qwen3-VL-4B and is included as an additional reference for allocating a similar total parameter budget to a larger shared backbone.

Results and Analysis. As shown in Table 4, HiUni-MoT-2B improves the OneIG score from 0.291 to 0.345 over HiUni-Dense-2B, while the average understanding score increases slightly from 58.55 to 59.23. This comparison suggests that task-specific backbone computation can benefit image generation under the same training setting. Nevertheless, HiUni-Dense-2B supports both acceptable capabilities with a fully shared backbone. We further include HiUni-Dense-4B as a parameter budget reference, which reaches 65.04 in understanding and 0.314 on OneIG. Under a similar total parameter, HiUni-Dense-4B provides substantially stronger understanding while retaining competitive generation performance. Figure 4 provides the training dynamics, shows that task-specific is not quite necessary for joint modeling. In other words, the dense backbone provides a simpler design that preserves strong understanding while supporting generation.

4.4 Analysis of Residual Visual Pathway

Residual Feature Source. The shared ViT and patch merger provide a common semantic representation for understanding and generation, but they also reduce fine-grained spatial information before the diffusion head. We therefore examine which feature source is more suitable for the residual pathway used in HiUni. Specifically, we compare three configurations: directly using pixel-level features before patch embedding (*Pixel Features*), injecting features from the ViT (*ViT Features*), or even removing the residual pathway (*No Gate*). All other architectural and training settings are kept fixed. As shown in Table 5, both residual sources improve image generation over *No Gate*. The average OneIG score increases from 0.267 to 0.284 with ViT features and further to 0.314 with pixel-level features. Compared with ViT features, pixel-level features provide an additional gain

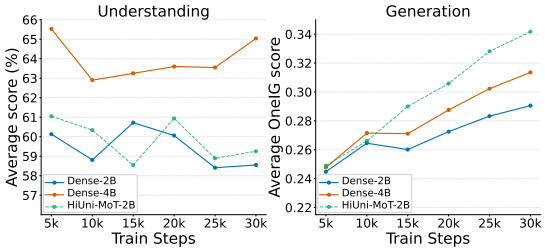


Figure 4: Training trajectories during the joint pretraining stage after warmup. We report the average understanding and OneIG scores here.

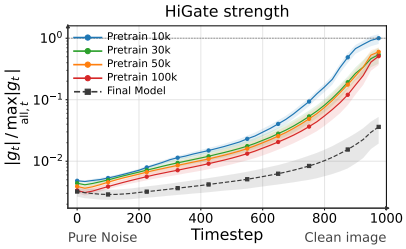


Figure 5: Evolution of *HiGate* injection strength across diffusion timesteps and training checkpoints.

Table 5: Analysis of *HiGate* with different insertion positions. Pixel Features and ViT Features use features from different stages of the visual pathway as the input to *HiGate*, while No Gate removes the gated residual pathway. Pixel Features corresponds to the default design of HiUni.

Configuration	Understanding							Generation
	$\text{MMBench}_{\text{EN}}$	RealWorldQA	MMStar	HallusionBench	AI2D	OCRBench	Avg.	OneIG Avg.
Pixel Features	72.60	65.49	53.93	48.96	77.95	71.30	65.04	**0.314**
ViT Features	74.61	67.45	53.07	49.05	77.82	69.00	65.17	0.284
No Gate	72.91	65.88	54.13	46.70	77.75	73.00	65.06	0.267

of 0.030, suggesting that features closer to the pixel space better complement the semantic representation produced by the shared ViT. Meanwhile, the average understanding score remains almost unchanged across the three settings, ranging from 65.04 to 65.17. These results support our choice of directly injecting pixel-level information into the generation pathway of HiUni.

Residual Injection Strength. We further analyze how HiUni uses the pixel-level residual pathway during image generation. Figure 5 visualizes the learned gate strength over diffusion timesteps and training checkpoints. For each timestep, we compute the maximum gate magnitude and normalize it by the global maximum over all plotted timesteps and checkpoints. The gate remains weak at high noise levels and gradually becomes stronger as the input approaches the clean image, suggesting that the residual pathway contributes mainly at later denoising stages where fine-grained spatial details are more important. The overall gate strength also decreases as pretraining progresses, and early checkpoints rely more on the residual pathway. Combined with the residual feature source analysis above, these results suggest that HiUni can retain semantic modeling in the shared backbone while supplying generation-specific spatial details near the output stage.

4.5 Cross-Task Interaction

Beyond architectural sharing, we further examine whether supervision from one visual task affects the others under the shared backbone. Specifically, adding image editing supervision improves the average understanding score from 62.64 to 64.49, while a slightly decreases of OneIG from 0.538 to 0.528. It indicates that task interaction naturally emerges under shared training, although its effect is not uniformly positive across capabilities. Detailed results are provided in Appendix E.

5 Conclusion

We present HiUni, a unified multimodal model that harmonizes visual understanding and generation within a shared visual space. HiUni reuses the language aligned ViT and dense Transformer of a pretrained VLM for shared semantic modeling, while supplying generation specific spatial details only near the output stage. Experiments show that strong understanding and generation can coexist within such a highly shared architecture, without requiring separate visual encoders or task specific backbones. Beyond HiUni itself, we believe this simple and scalable design provides a promising path to extend mature VLMs into more unified models with richer generative capabilities, broader tasks, and more interactive multimodal scenarios.

AI USE STATEMENT

Generative AI tools, including GPT-5.6 Sol, were used to assist with literature organization, language polishing, manuscript drafting, LaTeX formatting, and the preparation of tables and figures. The authors independently formulated the research problem, designed HIUNI, determined the training and evaluation protocols, conducted and interpreted the experiments, and made all scientific decisions. All AI-assisted content was reviewed, revised, and verified by the authors, who take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the HIUNI architecture, training objectives, optimization schedule, and experimental protocols in the main text and supplementary material. The supplementary material further includes model configurations, training recipes, ablation settings, and implementation details. Source code, model configurations, evaluation scripts, and checkpoints will be released upon acceptance. As the training corpus contains private and license-restricted data, the complete dataset cannot be redistributed. We will instead provide the sources of publicly available data, preprocessing procedures, sampling ratios, evaluation protocols, and hyperparameters to support reimplementation and further evaluation. Exact reproduction of the reported results may require access to the original training data.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

X AI. Grok-1.5 vision preview, 2024. URL `https://x.ai/blog/grok-1.5v`.

Jinbin Bai, Wei Chow, Ling Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Shuicheng Yan. Humanedit: A high-quality human-rewarded dataset for instruction-based image editing. *arXiv preprint arXiv:2412.04280*, 2024.

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

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, et al. Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer. *arXiv preprint arXiv:2605.11061*, 2026.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *Advances in Neural Information Processing Systems*, 38, 2026.

Guanzhou Chen, Erfei Cui, Changyao Tian, Danni Yang, Ganlin Yang, Yu Qiao, Hongsheng Li, Gen Luo, and Hongjie Zhang. Scaleedit-12m: Scaling open-source image editing data generation via multi-agent framework. *arXiv preprint arXiv:2603.20644*, 2026a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

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

Zhizhou Chen, Shanyan Guan, Zhanxin Gao, En Ci, Yanhao Ge, Wei Li, Zhenyu Zhang, Jian Yang, and Ying Tai. Vins-120k: Ultra high-resolution image editing with a large-scale dataset. *arXiv preprint arXiv:2605.23518*, 2026b.

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

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

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

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.

GLM-Image Team. Glm-image: Auto-regressive for dense-knowledge and high-fidelity image generation, 2026.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

Jiahao Hu, Tianxiong Zhong, Xuebo Wang, Boyuan Jiang, Xingye Tian, Fei Yang, Pengfei Wan, and Di Zhang. Vivid-10m: A dataset and baseline for versatile and interactive video local editing. *arXiv preprint arXiv:2411.15260*, 2024.

Dengyang Jiang, Ruoyi Du, Zhennan Chen, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Xiangpeng Yang, Huanqia Cai, Aiming Hao, Yuming Jiang, et al. An empirical study of training pixel-space text-to-image diffusion models. *arXiv preprint arXiv:2608.16887*, 2026.

Jiwon Kang, Heeji Yoon, Jaewoo Jung, Jaewon Min, Minkyeong Jeon, Biyeon Hwang, Sangwon Jung, and Seungryong Kim. Transferability between understanding and generation in unified multimodal models. *arXiv preprint arXiv:2607.04423*, 2026.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, 2024.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, et al. Qwen-image-bench: From generation to creation in text-to-image evaluation. *arXiv preprint arXiv:2605.28091*, 2026.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024a.

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, 2024b.

Zhiheng Liu, Weiming Ren, Xiaoke Huang, Shoufa Chen, Tianhong Li, Mengzhao Chen, Yatai Ji, Sen He, Jonas Schult, Belinda Zeng, et al. Tuna-2: Pixel embeddings beat vision encoders for multimodal understanding and generation. *arXiv preprint arXiv:2604.24763*, 2026a.

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *Advances in Neural Information Processing Systems*, 38:129274–129297, 2026.

Meituan LongCat Team. Longcat-image-edit. `https://huggingface.co/meituan-longcat/LongCat-Image-Edit`, 2026.

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

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, et al. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24838–24848. IEEE, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Qwen Team. Qwen-image-edit-2511. `https://huggingface.co/Qwen/Qwen-Image-Edit-2511`, 2025.

Qwen Team. Qwen-image 2.1. `https://github.com/QwenLM/Qwen-Image-2.1`, 2026.

Zihan Su, Hongyang Wei, Kangrui Cen, Yong Wang, Guanhua Chen, Chun Yuan, and Xiangxiang Chu. Generation enhances understanding in unified multimodal models via multi-representation generation. *arXiv preprint arXiv:2601.21406*, 2026.

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

Meituan LongCat Team, Bin Xiao, Chao Wang, Chengjiang Li, Chi Zhang, Chong Peng, Hang Yu, Hao Yang, Haonan Yan, Haoze Sun, et al. Longcat-next: Lexicalizing modalities as discrete tokens. *arXiv preprint arXiv:2603.27538*, 2026.

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

Dianyi Wang, Chaofan Ma, Feng Han, Size Wu, Wei Song, Yibin Wang, Zhixiong Zhang, Tianhang Wang, Siyuan Wang, Zhongyu Wei, et al. Unireason 1.0: A unified reasoning framework for world knowledge aligned image generation and editing. *arXiv preprint arXiv:2602.02437*, 2026a.

Haifeng Wang, Hua Wu, Tian Wu, Yu Sun, Jing Liu, Dianhai Yu, Yanjun Ma, Jingzhou He, Zhongjun He, Dou Hong, et al. Ernie 5.0 technical report. *arXiv preprint arXiv:2602.04705*, 2026b.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In *International Conference on Learning Representations*, volume 2025, pp. 259–271, 2025.

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

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977. IEEE, 2025b.

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 e-prints*, pp. arXiv–2506, 2025c.

Penghao Wu, Haiwen Diao, Weichen Fan, Lewei Lu, Dahua Lin, and Ziwei Liu. Uncovering understanding-generation synergy in native unified multimodal models: From representation, task to system. *arXiv preprint arXiv:2609.01607*, 2026.

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. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17739–17750. IEEE, 2025d.

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

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

Jinbo Yan, Limeng Qiao, Jie Qin, Junyan He, Feize Wu, and Guanglu Wan. Unispace: Unified visual representation and scalable multimodal modeling. *arXiv preprint arXiv:2608.08676*, 2026.

Yang Ye, Xianyi He, Zongjian Li, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, Li Yuan, et al. Imgedit: A unified image editing dataset and benchmark. *Advances in Neural Information Processing Systems*, 38, 2026.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. Pixeldit: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14273–14282, 2026.

Guozhen Zhang, Xuerui Qiu, Yutao Cui, Tianhui Song, Changlin Li, Junzhe Li, Tao Huang, Xiao Zhang, Yang Li, Jianbing Wu, et al. Hydra-x: Native unified multimodal models with holistic visual tokenizers. *arXiv preprint arXiv:2606.13289*, 2026a.

Yichen Zhang, Da Peng, Zonghao Guo, Zijian Zhang, Xuesong Yang, Tong Sun, Shichu Sun, Yidan Zhang, Yanghao Li, Haiyan Zhao, et al. Cheers: Decoupling patch details from semantic representations enables unified multimodal comprehension and generation. *arXiv preprint arXiv:2603.12793*, 2026b.

Dian Zheng, Manyuan Zhang, Hongyu Li, Kai Zou, Hongbo Liu, Ziyu Guo, Kaituo Feng, Yexin Liu, Ying Luo, and Hongsheng Li. Aia: Rethinking architecture decoupling strategy in unified multimodal model. *arXiv preprint arXiv:2511.22663*, 2025.

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. In *International Conference on Learning Representations*, volume 2025, pp. 6446–6469, 2025.

Table 6: Training recipe of HIUNI. Image resolutions denote the maximum resolution of an image, subject to the image number and sequence length constraint. "-" denotes an inapplicable value.

	Generation Warmup	Pretraining				SFT	
		Phase I	Phase II	Phase III	Phase IV	Phase I	Phase II
Hyperparameters							
Peak learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	2×10^{-5}	2×10^{-5}	2×10^{-5}
Min learning rate	-	-	-	2×10^{-5}	-	-	2×10^{-6}
LR scheduler	Constant	Constant	Constant	Cosine decay	Constant	Constant	Cosine decay
Optimizer	AdamW ($\beta_1 = 0.9,\ \beta_2 = 0.95,\ \epsilon = 10^{-8}$)						
Batch size	256	512	512	512	1024	1024	2048
Training steps	30K	115K	105K	60K	30K	14K	8K
Warmup steps	–	2000	2000	2000	2000	2000	2000
Loss weight ($\mathcal{L}_{\text{txt}}$:$\mathcal{L}_{\text{gen}}$)	0 : 1	0.1 : 1	0.1 : 1	0.1 : 1	0.1 : 1	0.1 : 1	0.1 : 1
Sequence length	8192						
Generation resolution	256^2	$256^2 \rightarrow 512^2$	$512^2 \rightarrow 1024^2$	$512^2 \rightarrow 1024^2$	$1024^2 \rightarrow 2048^2$	$1024^2 \rightarrow 2048^2$	$1024^2 \rightarrow 2048^2$
Understanding resolution	–	$512^2 \rightarrow 2048^2$	$512^2 \rightarrow 2048^2$	$512^2 \rightarrow 2048^2$	$512^2 \rightarrow 2048^2$	$512^2 \rightarrow 2048^2$	$512^2 \rightarrow 2048^2$
Task sampling ratio							
Pure-text modeling	-	0.10	0.10	0.10	0.10	0.05	0.05
Image understanding	-	0.30	0.30	0.30	0.30	0.15	0.20
Image generation	1.00	0.60	0.60	0.60	0.60	0.60	0.50
Image editing	-	-	-	-	-	0.20	0.25

Table 7: Configuration of the *Pixel Diffusion Head* and *HiGate*.

Configuration	
Number of Transformer blocks	10
Hidden dimension	$D = 2560$
Attention	GQA with 32 query heads and 8 key-value heads
Head dimension	80
Feed-forward network	SwiGLU, intermediate dimension 9728
Flattened pixel dimension	$3 \times 32 \times 32 = 3072$
Pixel residual projection	$3072 \rightarrow 2560$
Semantic input	Diffusion-head hidden state $\mathbf{h} \in \mathbb{R}^{2560}$
Pixel input	Projected pixel residual $\mathbf{u} \in \mathbb{R}^{2560}$
Gate MLP	$2560 \rightarrow 2560 \rightarrow 1$ with SiLU activation
Block-level AdaLN	Disabled

A TRAINING DETAILS

This section provides additional training details of HIUNI. Following the three-stage procedure described in Section 4.1, training consists of generation warmup, joint pretraining, and supervised fine tuning. The complete training configuration is summarized in Table 6.

A.1 TRAINING STAGES

During generation warmup, we train the *Pixel Diffusion Head* and *HiGate* for 30K steps while keeping the pretrained visual encoder and backbone frozen. Only image generation data are used in this stage, with a fixed resolution of 256^2.

We then unfreeze the full model and perform joint pretraining for 310K steps on pure text modeling, image understanding, and image generation. The generation resolution is progressively increased from 256^2 to 2048^2, while the task sampling ratio remains $0.10{:}0.30{:}0.60$ throughout pretraining.

Finally, supervised fine tuning is performed for 22K steps. Image editing is introduced at this stage while the existing tasks are retained. The task mixture is adjusted between the two fine tuning phases as shown in Table 6. Other phase specific settings, including batch size, learning rate schedule, and image resolution, are also provided in the table.

A.2 PIXEL DIFFUSION HEAD CONFIGURATION

The *Pixel Diffusion Head* contains ten Transformer blocks with a hidden dimension of 2560, matching the hidden dimension of the shared backbone. For each target visual token, the corresponding 32×32 RGB pixel patch is flattened and projected to the diffusion hidden dimension before being injected through *HiGate*. The gate is predicted by a two layer MLP with SiLU activation. We do not apply an additional sigmoid or tanh activation to the gate output.

A.3 Optimization and Diffusion Settings

We use AdamW throughout training with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 10^{-8}$. The maximum sequence length is 8192 tokens. For image editing, the reference and target images share the sequence budget with textual and special tokens, and their resolutions are adjusted jointly to keep the complete sequence within 8192 tokens. The relative loss weight between language modeling and diffusion is set to $0.1{:}1$, except for warmup.

For diffusion training, timesteps are sampled from a logit normal distribution with $\mu = -0.8$ and $\sigma = 0.8$. We use a resolution dependent noise scale according to the number of spatial tokens N_g. Specifically, we sample $\mathbf{z} \sim \mathcal{N}(0, I)$ and define

$$\boldsymbol{\epsilon}_{N_g} = \sigma_{N_g}\mathbf{z}, \qquad \sigma_{N_g} = \sigma_0\sqrt{\frac{N_g}{N_0}},$$

where $\sigma_0 = 1$, $N_0 = 64$, and σ_{N_g} is clipped to $[1, 8]$. The noisy input and target velocity are

$$\mathbf{x}_t = t\mathbf{x}_0 + (1-t)\boldsymbol{\epsilon}_{N_g}, \qquad \mathbf{v} = \mathbf{x}_0 - \boldsymbol{\epsilon}_{N_g}.$$

Since the diffusion head directly predicts the clean image $\hat{\mathbf{x}}_0$, we convert it to the velocity prediction used by the training objective as

$$\hat{\mathbf{v}} = \frac{\hat{\mathbf{x}}_0 - \mathbf{x}_t}{\max(1-t, 0.05)}.$$

The denominator is clipped only for numerical stability when t approaches one and does not modify the forward interpolation or the target velocity.

The maximum multimodal sequence length is 8192 tokens. With an effective visual patch size of 32×32, an image of size $H \times W$ contributes approximately $(H/32)(W/32)$ visual tokens.

A.4 Loss Aggregation

Language modeling and diffusion losses are normalized separately. For pure text modeling and image understanding, cross entropy is averaged over all valid textual targets:

$$\mathcal{L}_{\text{txt}} = -\frac{1}{|\mathcal{T}|}\sum_{(n,j)\in\mathcal{T}} \log p_\theta\left(y_{n,j} \mid y_{n,<j}, \mathbf{c}_n\right),$$

where $\mathcal{T}$ denotes all valid textual targets in the batch.

For image generation and editing, we first compute the mean squared velocity error independently for each image:

$$\ell_{\text{diff}}^{(n)} = \frac{1}{D_n}\left\|\hat{\mathbf{v}}_n - \mathbf{v}_n\right\|_2^2,$$

where D_n denotes the number of elements in the velocity target, including all spatial positions and channels. The diffusion objective is then averaged over all generation and editing examples:

$$\mathcal{L}_{\text{gen}} = \frac{1}{|\mathcal{I}_{\text{gen}}|}\sum_{n\in\mathcal{I}_{\text{gen}}} \ell_{\text{diff}}^{(n)},$$

where $\mathcal{I}_{\text{gen}}$ contains both text to image generation and image editing examples. This per image normalization prevents higher resolution samples from receiving larger effective weights, while the relative contribution of generation and editing follows the task sampling ratios in Table 6.

A.5 Training Dataset

Figure 3 summarizes the composition of the training corpus used for image understanding, generation, and editing. For image understanding, the data are mainly distributed across OCR and visual grounding, together with smaller portions of arts and sciences and agent-related data. For image generation, the corpus is organized into three major groups, namely nature, people, and design. For image editing, the data mainly cover subject-level, attribute-level, and structure-level edits.

The image understanding and generation corpora are constructed from proprietary datasets. The image editing corpus contains about 4M examples collected from public datasets and prior work (Ye

et al., 2026; Wei et al., 2025; Bai et al., 2024; Chen et al., 2026a;b; Wu et al., 2025c; Hu et al., 2024), together with additional private editing pairs produced by our internal data-generation pipeline. For ImgEdit, only the official training split is included in the training corpus, while evaluation is conducted exclusively on its official test split.

B Evaluation Details

B.1 Image Understanding

We evaluate HiUni on the image understanding benchmarks reported in Table 1. We follow the official benchmark splits and scoring protocols, with benchmark specific answer extraction when required. For the standard image understanding benchmarks, we use `min_pixels` $= 262144$, `max_pixels` $= 4194304$, and `max_model_len` $= 16384$. For evaluations with reasoning mode, the model first generates its reasoning process and then produces the final answer. Only the final answer is passed to the benchmark scorer. If a valid answer is not completed in the first generation, decoding is continued from the existing context without restarting the example.

We follow the official benchmark splits and scoring protocols for all reported evaluations.

MathVista. We evaluate on the official MathVista-Mini split and follow its standard evaluation protocol and scorer. The reported score is averaged over two complete evaluations.

MMBench. We evaluate MMBench-DEV-EN using the official circular evaluation protocol and report the corresponding overall score.

RealWorldQA. We evaluate the official RealWorldQA set and report the EM Overall score.

MMStar. We evaluate the official MMStar set and report the Overall score across all categories.

HallusionBench. We follow the official evaluation protocol and report its Overall score.

AI2D. We evaluate on AI2D-Test and use the official EM protocol for multiple choice prediction.

OCRBench. We evaluate OCRBench v1 and report the official Final Score as a percentage.

OmniDocBench. We evaluate the English split of OmniDocBench v1.0 and report the official OverallEdit score, where a lower value indicates better performance.

B.2 Image Generation

We use the original benchmark prompts without prompt rewriting or post processing. The score is computed by the official evaluators, and the inference settings for are summarized below.

QwenImageBench. We use a CFG scale of 6, a shift of 5, and 100 sampling steps, and generate one image for each prompt.

OneIG. We use a CFG scale of 6, a shift of 4, and 25 sampling steps, and generate four images for each prompt.

LongText. We use a CFG scale of 8, a shift of 8, and 100 sampling steps, and generate four images for each prompt.

B.3 Image Editing

We evaluate HiUni on ImgEdit and GEdit using the same image editing inference pipeline. For both benchmarks, we generate one image for each instruction using 50 sampling steps with the Euler scheduler, a shift of 3.0, and a CFG scale of 4.0. No additional post processing is applied.

Table 8: Detailed results on OneIG.

Models	Size	OneIG-ZH						OneIG-EN					
		Alignment	Text	Reasoning	Style	Diversity	Overall	Alignment	Text	Reasoning	Style	Diversity	Overall
Generation-only Models													
GLM-Image*	9B + 7B	0.738	0.976	0.284	0.335	0.221	0.511	0.805	0.969	0.298	0.353	0.213	0.528
Qwen-Image*	20B	0.825	0.963	0.267	0.405	0.279	0.548	0.882	0.891	0.306	0.418	0.197	0.539
ERNIE-Image*	8B	0.842	0.898	0.266	0.421	0.177	0.521	0.891	0.967	0.295	0.447	0.169	0.554
Unified Multimodal Models													
Bagel	7B + 7B	0.672	0.365	0.186	0.357	**0.268**	0.370	0.769	0.244	0.173	0.367	**0.251**	0.361
InternVL-U*	2B + 1.7B	0.750	0.900	0.230	0.370	0.260	0.500	0.820	0.740	0.270	0.400	0.250	0.500
SenseNova-8B*	8B + 8B	**0.826**	**0.977**	**0.303**	0.392	0.176	**0.535**	**0.882**	**0.969**	**0.330**	0.396	0.166	0.549
UniSpace*	8B + 8B	0.807	0.881	0.276	**0.455**	0.244	0.533	0.860	0.937	0.311	**0.467**	0.233	**0.561**
HiUni	4B + 1.3B	0.795	0.953	0.259	0.426	0.185	0.524	0.851	0.941	0.275	0.443	0.179	0.538

Table 9: Detailed results on GEdit.

Models	Size	Chinese prompts											
		background	color	material	motion	human	style	add	remove	replace	text	tone	**Avg.**
Editing-only Models													
Qwen-Image-2.1	7B	6.83	7.26	6.75	6.73	6.04	6.88	7.51	4.29	6.87	7.66	7.22	6.73
LongCat-Image-Edit	6B	7.25	7.41	6.97	7.20	6.39	6.80	7.53	7.51	7.34	7.32	6.92	7.15
Qwen-Image-Edit-2511	20B	7.01	7.59	6.79	7.23	6.66	6.85	7.54	7.48	7.28	7.65	7.00	7.19
Unified Multimodal Models													
InternVL-U	2B + 1.7B	6.55	5.66	5.32	5.92	5.42	6.15	6.12	5.31	6.25	4.55	5.87	5.74
Bagel	7B + 7B	6.26	6.59	5.97	6.26	5.44	5.83	7.05	6.99	6.83	6.30	5.91	6.31
SenseNova-U1	8B + 8B	**7.01**	**7.28**	**6.95**	**7.23**	**6.33**	**6.81**	**7.39**	**7.17**	7.01	**7.53**	**6.89**	**7.05**
UniSpace	8B + 8B	6.62	7.00	6.40	6.84	6.05	6.73	5.77	7.15	**7.03**	6.26	6.44	6.57
HiUni	4B + 1.3B	6.65	7.03	6.40	6.80	6.01	6.42	7.08	6.92	6.92	7.28	6.18	6.70

Models	Size	English prompts											
		background	color	material	motion	human	style	add	remove	replace	text	tone	**Avg.**
Editing-only Models													
Qwen-Image-2.1	7B	7.08	7.40	6.82	7.00	6.25	6.57	7.43	5.31	7.25	7.82	7.24	6.92
LongCat-Image-Edit	6B	7.20	7.43	6.72	7.20	6.28	6.86	7.44	7.34	7.20	7.37	7.02	7.10
Qwen-Image-Edit-2511	20B	7.12	7.53	6.85	7.42	6.53	6.93	7.47	7.34	7.19	7.73	6.80	7.17
Unified Multimodal Models													
InternVL-U	2B + 1.7B	6.29	5.86	5.44	6.61	5.21	6.35	6.32	5.03	6.18	4.65	5.25	5.74
Bagel	7B + 7B	6.28	6.49	5.63	6.51	5.32	5.89	7.03	6.67	6.87	6.48	5.77	6.27
SenseNova-U1	8B + 8B	**7.00**	6.94	6.59	**7.05**	6.05	**6.67**	**7.28**	**7.25**	7.04	**7.55**	**6.86**	**6.94**
UniSpace	8B + 8B	6.63	6.78	**6.63**	**7.05**	**6.17**	6.57	6.61	6.88	**7.16**	6.30	6.70	6.68
HiUni	4B + 1.3B	6.80	**7.14**	6.43	6.98	6.10	6.58	7.16	7.02	6.98	7.21	6.57	6.81

ImgEdit. We follow the official evaluation protocol and prompts, using GPT-5.2 as the judge model. Each edited image is scored on a 0 to 5 scale, and the Overall score is averaged across edit types.

GEdit. We follow the VIEScore evaluation protocol and use GPT-5.2 as the judge model. Each result is evaluated on semantic consistency (G_SC) and perceptual quality (G_PQ), both on a 0 to 10 scale, with their geometric mean reported as G_O. Scores are first averaged within each edit type and then across all edit types.

C Optimization Stability of the Pixel Diffusion Head

During early development, we observed that the *Pixel Diffusion Head* with AdaLN frequently produced abrupt loss spikes when trained directly in pixel space. We therefore simplified the diffusion blocks by removing repeated AdaLN modulation while retaining timestep conditioning through the shared backbone and the final prediction layer. Figure 6 shows representative training trajectories before and after this modification. The run without AdaLN exhibits a noticeably smoother optimization trajectory with fewer large fluctuations, consistent with the training behavior observed in our subsequent experiments. Since the two runs were obtained at different development stages and use different resolution schedules, we use them only as a qualitative comparison of optimization stability rather than a controlled comparison of absolute diffusion losses.

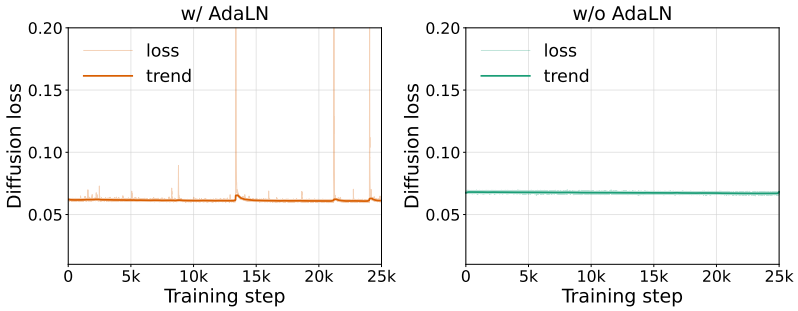


Figure 6: Diffusion training trajectories from development runs with and without AdaLN in the Pixel Diffusion Head. The run without AdaLN (right) shows fewer abrupt loss spikes. The two curves were collected at different development stages and therefore provide only a qualitative comparison of optimization behavior rather than a controlled isolation of the effect of AdaLN.

Table 10: Results on the WISE benchmark for chain of thought image generation.

Model	Overall	Cultural	Time	Space	Biology	Physics	Chemistry
DeepGen 1.0*	0.57	0.65	0.41	0.72	0.39	0.59	0.45
Bagel$_{\mathrm{w/cot}}{}^\star$	0.63	0.78	0.63	0.57	0.38	0.55	0.51
SenseNova-U1-8B*	0.68	0.69	0.60	0.74	0.63	0.72	0.69
SenseNova-U1-8B$_{\mathrm{w/cot}}{}^\star$	0.79	0.80	0.74	0.86	0.76	0.78	0.80
Qwen-Image-2.0*	0.80	0.82	0.65	0.90	0.79	0.80	0.75
GPT-Image-1.5*	0.83	0.89	0.69	0.88	0.80	0.76	0.78
HiUni$_{\mathrm{w/cot}}$	0.75	0.81	0.68	0.78	0.63	0.74	0.78

D Extending to Chain-of-Thought Image Generation

Formulation As an additional experiment, we extend HiUni to chain-of-thought (CoT) based image generation, where the model first produces an intermediate textual reasoning sequence before generating the target image. Given a text prompt $\mathbf{p}$, the model autoregressively generates a textual reasoning sequence $\mathbf{r}$, which is then included in the conditioning context as $\mathbf{C}_{\mathrm{cot}} = (\mathbf{p}, \mathbf{r})$. Image generation follows the same formulation as in Section 3.3, $\hat{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t, t, \mathbf{C}_{\mathrm{cot}})$. The reasoning sequence is optimized with the language modeling objective $\mathcal{L}_{\mathrm{txt}}$, while the target image is optimized with the same diffusion objective $\mathcal{L}_{\mathrm{gen}}$ used for standard image generation.

Training We obtain the CoT model by continuing training from a HiUni checkpoint after Phase III of joint pretraining, where the training data are derived from UniReason (Wang et al., 2026a). The model is further trained for 10K steps with a global batch size of 512. Within the image generation data, standard text-to-image examples and CoT generation examples are sampled with a ratio of 1:1, while image understanding and pure-text modeling data are retained during training. No additional backbone or generation module is introduced.

Results We evaluate the extension on the WISE benchmark (Niu et al., 2025), with results reported in Table 10. HiUni$_{\mathrm{w/cot}}$ achieves an Overall score of 0.75, exceeding DeepGen 1.0 and Bagel with CoT and remaining competitive with stronger recent image generation models. It performs particularly well on Cultural, Time, and Chemistry, with scores of 0.81, 0.68, and 0.78, respectively. These results are obtained through a short fine-tuning stage without modifying the shared backbone or generation pathway, indicating that the HiUni architecture can be readily extended to reasoning based image generation. In future work, we plan to integrate this capability into the main training pipeline and further study its interaction with standard image generation.

Table 11: Effect of image-editing supervision on image understanding and generation. All other settings are kept fixed, with image-editing data as the only changed factor.

Configuration	Understanding							Generation
	$\text{MMBench}_{\text{EN}}$	RealWorldQA	MMStar	HallusionBench	AI2D	OCRBench	Avg.	OneI Avg.
w/o Editing Supervision	69.58	64.58	50.20	48.43	72.54	70.50	62.64	0.538
w/ Editing Supervision	72.29	66.80	55.00	43.83	79.24	69.80	64.49	0.528

E Effect of Editing Supervision

We further examine cross-task interaction by comparing two training settings that differ only in the inclusion of image-editing supervision. The understanding and generation data, model architecture, and optimization settings are kept unchanged.

As shown in Table 11, editing supervision improves the average understanding score from 62.64 to 64.49, with gains on four of the six evaluated benchmarks. The improvements are particularly pronounced on AI2D-Test and MMStar, which increase by 6.70 and 4.80 points, respectively, which indicate that the transferred supervision may be particularly useful for structured visual reasoning. In contrast, HallusionBench decreases by 4.60 points, while OCRBench changes only slightly. This non-uniform pattern suggests that the benefit of editing supervision is task dependent rather than a general improvement across all understanding capabilities. At the same time, the average OneIG score slightly decreases from 0.538 to 0.528. Together with the understanding results, this suggests that editing supervision introduces both positive transfer and mild interference under shared training. It can improve several understanding capabilities, while also changing the optimization balance among tasks rather than uniformly benefiting all of them.

F Qualitative Results

F.1 Image Generation

Representative image generation results of HiUni are shown in Figure 7, and text rendering results of HiUni are shown in Figure 8.

F.2 Image Editing

Representative image editing results of HiUni are shown in Figure 9, covering several common editing types.

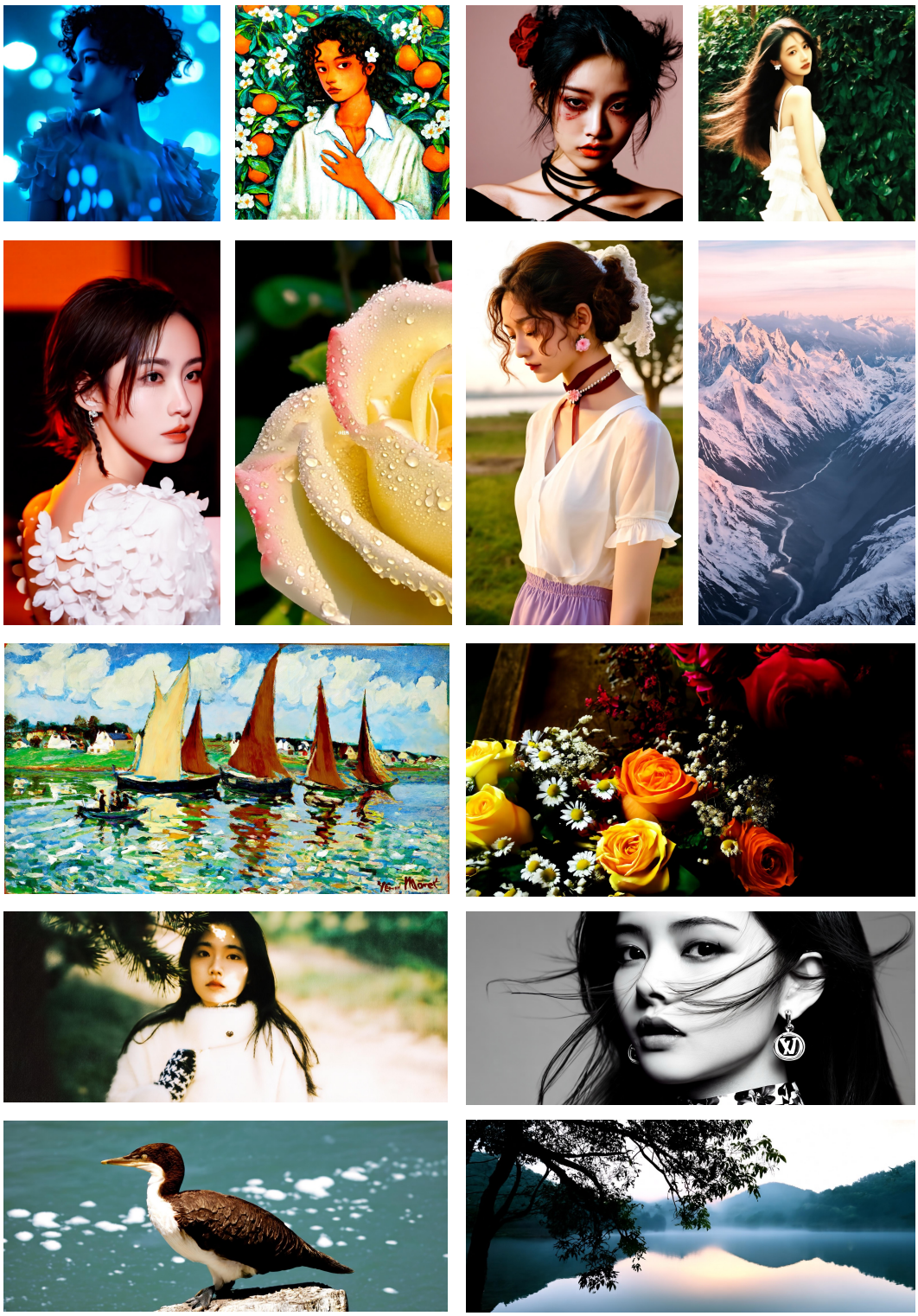


Figure 7: Showcase of HIUNI in image generation.

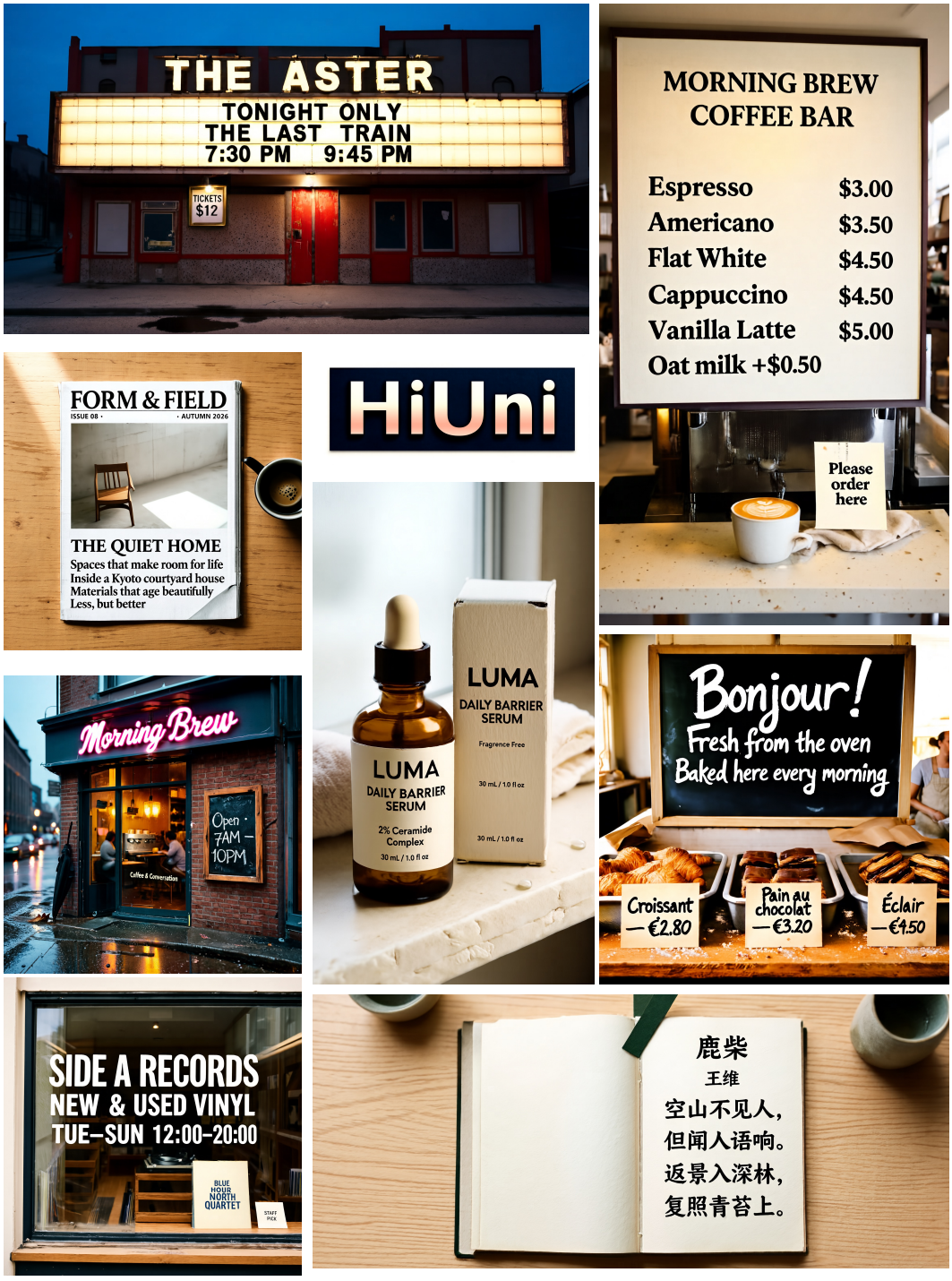


Figure 8: Showcase of HiUni in text rendering.



Figure 9: Showcase of HIUNI in image editing.