

HBridge: H-Shape Bridging of Heterogeneous Experts for Unified Multimodal Understanding and Generation

Xiang Wang¹ Zhifei Zhang² He Zhang² Zhe Lin² Yuqian Zhou² Qing Liu² Shiwei Zhang¹ Yijun Li²
Shaoteng Liu² Haitian Zheng² Jason Kuen² Yuehuan Wang¹ Changxin Gao¹ Nong Sang¹

¹Key Laboratory of Image Processing and Intelligent Control,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
²Adobe Research

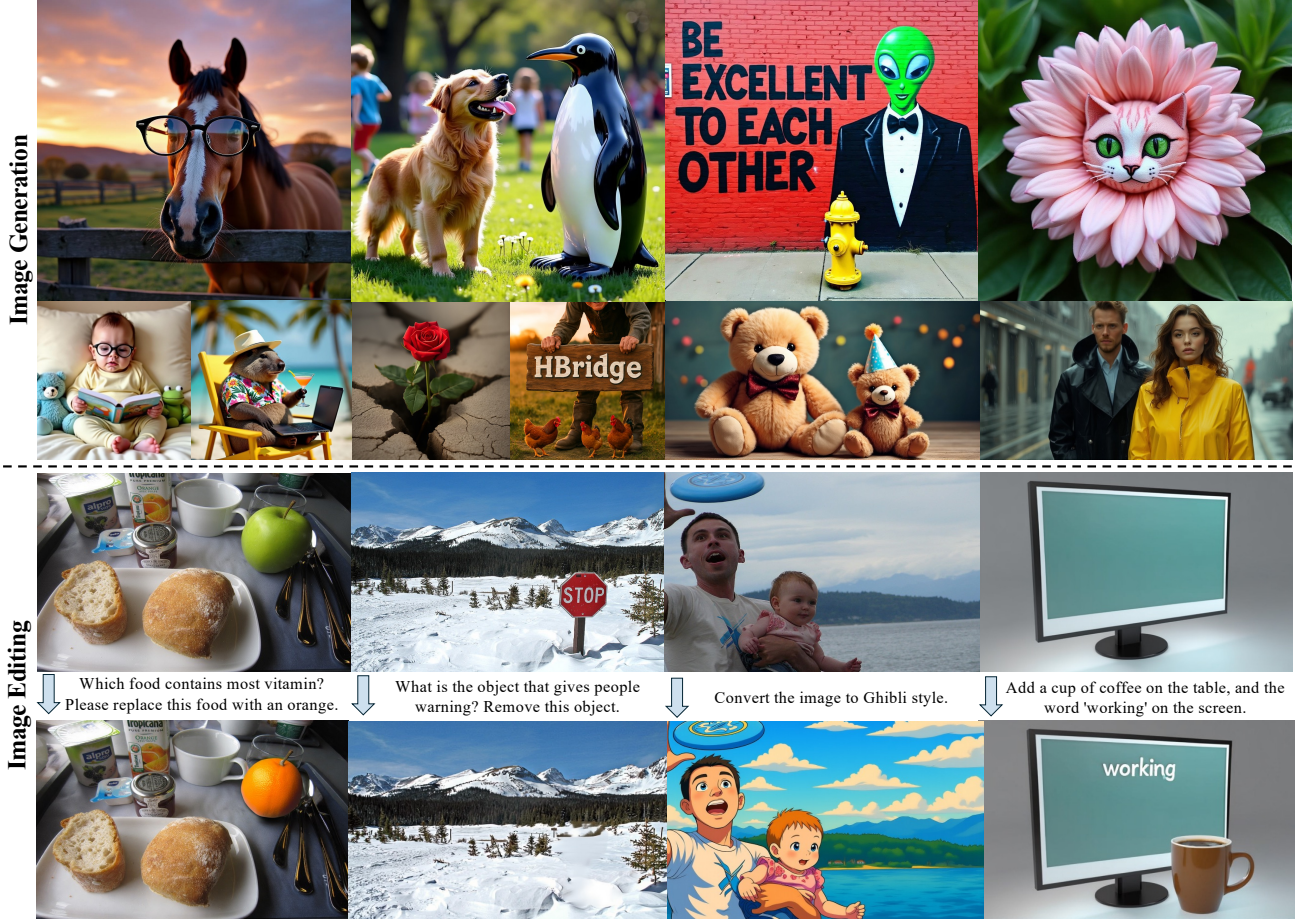


Figure 1. Image generation and editing samples from HBridge, which achieves high-quality and photorealistic results.

Abstract

Recent unified models integrate understanding experts (e.g., LLMs) with generative experts (e.g., diffusion models), achieving strong multimodal performance. However, recent advanced methods such as BAGEL and LMFusion follow the Mixture-of-Transformers (MoT) paradigm, adopting a symmetric design that mirrors one expert to another for convenient initialization and fusion, which

remains suboptimal due to inherent modality discrepancies. In this work, we propose HBridge, an asymmetric H-shaped architecture that enables heterogeneous experts to optimally leverage pretrained priors from their respective modality domains. Unlike prior dense fusion strategies that straightforwardly connect all layers between experts via shared attention, HBridge selectively bridges intermediate layers, reducing over 40% attention sharing, which improves efficiency and enhances generation quality. Shallow

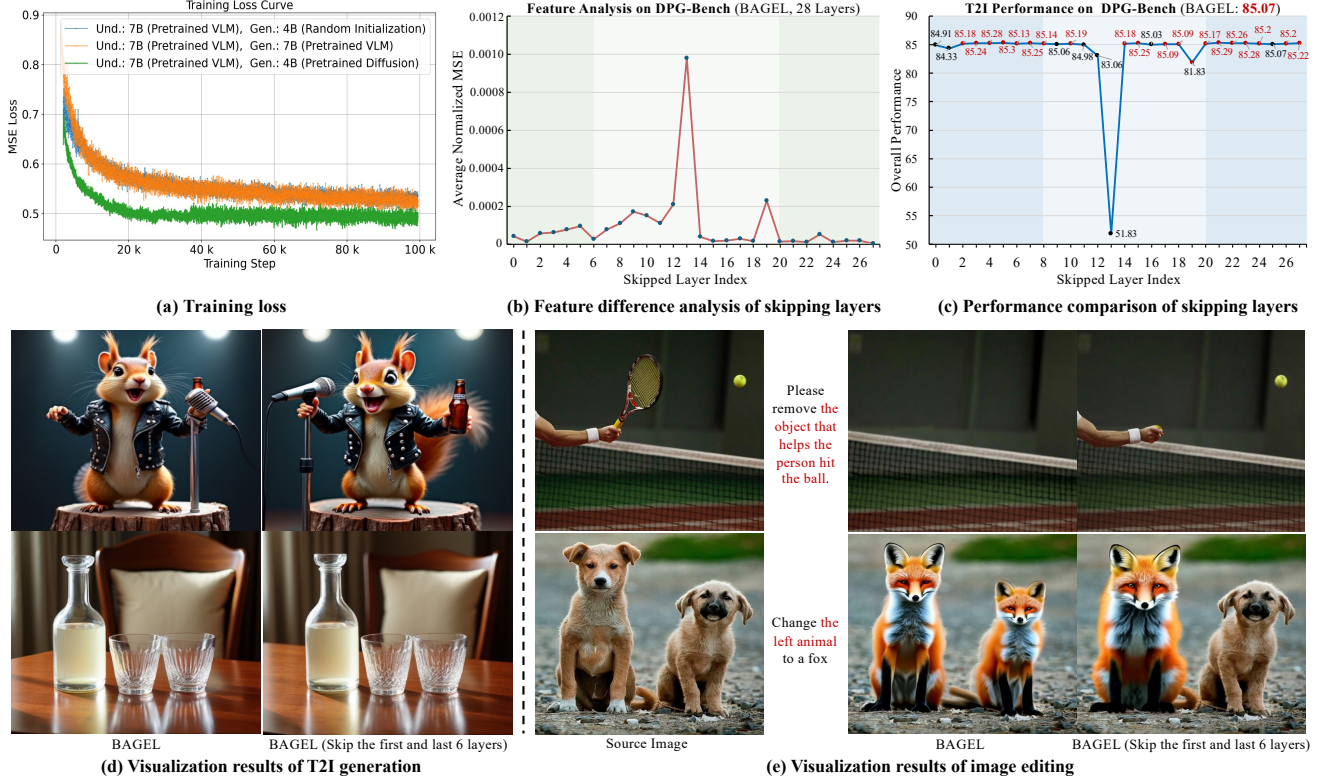


Figure 2. Motivation behind HBridge. The “H” refers to Heterogeneous (asymmetric) experts and the H-shape bridging that connects only the mid-layers between experts. Our asymmetric design is motivated by its faster convergence, as illustrated in (a). The symmetric 7B+7B baseline (orange), initialized from Qwen2.5-VL-7B, converges slower than our asymmetric 7B+4B model (green), where the 4B generative expert is initialized from OmniGen2. Remarkably, even a randomly initialized 7B+4B model (blue) converges at a similar speed to the well initialized 7B+7B model. The H-shape bridging is inspired by the correlation patterns in (b) and (c), which measure feature drift and T2I performance degradation when breaking individual cross-expert self-attention connections in BAGEL. These reveal that mid-layer connections dominate the performance, while shallow and deep-layer connections contribute minimally. By retaining only the essential mid-layer bridges, our architecture not only simplifies fusion but can outperform the original dense-fusion model. The corresponding visual examples in (d) and (e) further validate the above observation. Please note that results in (b)-(e) are derived directly from BAGEL.

and deep layers, which capture modality-specific representations, are decoupled, while mid-layer bridging promotes semantic alignment. To further strengthen cross-modal coherence, we introduce semantic reconstruction tokens that explicitly guide the generative expert to reconstruct visual semantic tokens of the target image. Extensive experiments across multiple benchmarks demonstrate the effectiveness and superior performance of HBridge, establishing a new paradigm for unified multimodal generation.

1. Introduction

Unified multimodal understanding and generation [5, 7, 8, 10, 18, 19, 26, 33, 36, 45, 46, 49, 54–61, 63, 65, 70, 73, 76, 77] has recently gained significant attention and emerged as a central direction for bridging understanding and generation within a unified architecture, powering multifunctional applications. Current methods such as BAGEL [10], BLIP3o [7], MetaQuery [37], Mogao [28], OmniGen2 [58], and LMFusion [45] exemplify this trend.

They usually combine autoregressive large (vision-) language models (LLMs/VLMs) [2, 50] with a diffusion-based visual generator [12, 43], where LLMs/VLMs are used for understanding tasks, and the diffusion generator focuses on visual synthesis. Among these methods, the Mixture-of-Transformers (MoT) paradigm [10, 28, 45] employed by BAGEL and LMFusion demonstrates impressive capabilities and achieves state-of-the-art results across both understanding and generation tasks. This paradigm uses a symmetric, densely connected design: it deploys two identical experts, both initialized from a pretrained LLM/VLM for understanding and generation, and bridges multimodal interactions through layer-by-layer shared self-attention.

Despite rapid progress, the unified MoT paradigm faces two fundamental limitations:

1) **Symmetric architectures restrict generative priors.** The generative branch is typically initialized from an autoregressive LLM, since large-scale pretrained diffusion backbones with LLM-compatible architectures are not available.

This mismatch prevents the generative expert from benefiting from strong pretrained priors, resulting in slow convergence (Fig. 2(a)) and high training cost, often comparable to random initialization. In addition, understanding and generative models follow divergent scaling trends: language models now exceed 1T parameters [19, 23, 47], while state-of-the-art generative models remain mostly below 20B [12, 22]. These discrepancies indicate that heterogeneous experts are needed to fully exploit the strengths of existing pretrained models.

2) **Dense layer-wise attention sharing ignores task asymmetry.** Understanding tasks usually rely on high-level semantic reasoning, whereas generation requires modeling fine-grained low-level structures. Sharing multimodal self-attention across all layers, including the earliest input layers and final output layers, can interfere with learning the task-specific feature spaces required by each branch. Empirically, as shown in Fig. 2 (b-e), the early and late layers of BAGEL [10] contribute minimally to final performance; skipping these connections does not degrade results and can even improve them. Moreover, our analysis shows that dense layer-by-layer connections may encourage the generative expert to overfit shallow features from the understanding expert (Fig. 6), reducing its ability to capture high-level contextual semantics (Fig. 5).

Motivated by these observations, we propose **HBridge**, an asymmetric H-shaped MoT architecture that unifies multimodal understanding and generation through heterogeneous experts and a mid-layer semantic bridge:

- Unlike prior symmetric designs, **HBridge** pairs a large pretrained LLM with a diffusion-based generative expert, and replaces full-layer sharing with a selective mid-layer bridge, eliminating over 40% of attention connections and preventing potential shallow overfitting.
- We introduce semantic reconstruction tokens that guide the generative branch to reconstruct ViT-level visual features, further enhancing semantic reasoning.
- Extensive experiments demonstrate that **HBridge** offers strong resource efficiency, attaining superior performance under lower training budgets. Remarkably, compared to BAGEL’s ~ 2.5 T T2I tokens, **HBridge** requires ~ 200 B T2I training tokens yet achieves even higher performance.

2. Related Work

Image Generation and Editing. Diffusion models [1, 3, 4, 6, 9, 12, 31, 35, 38, 39, 41–44, 66, 72, 75] have become the dominant framework for plausible image generation and controllable editing. Stable Diffusion [43] introduces latent diffusion models, operating in a compressed latent space to achieve high-quality, efficient text-to-image synthesis. Recent advances in diffusion Transformers (DiT) [38] have scaled visual generation to a new level. PixArt- α [6] leverages cross-attention modules to inject text conditions

into DiT backbone for large-scale text-to-image training. SD3 [12] and FLUX [22] introduce hybrid multimodal DiT blocks with cross-modal conditioning and apply flow matching [30] to optimize the model, significantly improving compositional semantic control and visual realism. For image editing, typical methods such as InstructPix2Pix [4], ControlNet [72] and Composer [17] enable text-guided or structural control by fine-tuning conditional branches or adding spatial customized adapters.

Unified Understanding and Generation. Unified understanding and generation models [5, 7, 10, 28, 29, 45, 54] aim to jointly perform perception and generation within one framework. Two major paradigms have emerged: pure autoregressive (AR) and hybrid AR/diffusion architectures. Pure autoregressive approaches [20, 25, 41, 46] treat multimodal data as token sequences, leveraging language modeling objectives for both understanding and generation. Chameleon [46] integrates discrete visual tokens into a uniform Transformer-based architecture and trains the whole model from scratch in an end-to-end manner. Uni-Token [20] incorporates a combination of discrete and continuous tokens to encode visual features and trains an autoregressive generation model for unified tasks. However, recent AR models still struggle to achieve photorealistic synthesis due to the difficulty of unified tokenizers and the accumulation errors of autoregressive prediction. The hybrid AR/diffusion paradigm [5, 10, 28, 34, 37, 45, 49, 52, 54, 63, 64, 76] combines autoregressive LLM and diffusion model, and has attracted widespread attention. Meta-Query [37] optimizes a set of learnable queries on the input side of the LLM and feeds the corresponding outputs into the diffusion model for generation. Metamorph [49] learns an LLM to predict continuous visual tokens, guiding the diffusion generator to synthesize visual content. Recently, BAGEL [10] and Mogao [28] introduced a dense Mixture-of-Transformers (MoT) architecture that bridges the understanding and generative experts of identical Transformer architecture via fully shared attention, achieving state-of-the-art visual perception and synthesis performance. While effective, they usually use symmetric experts with pretrained LLM initialization, overlooking the enormous potential of exploring powerful pretrained diffusion models to reduce resource consumption. In contrast, **HBridge** departs from these symmetric paradigms by introducing asymmetric experts with mid-layer semantic bridge, enabling the utilization of pretrained diffusion weights and preserving task-specific representation learning.

3. HBridge

We propose **HBridge**, a hybrid asymmetric MoT architecture that unifies multimodal understanding and generation within a single framework. Without loss of generality, we

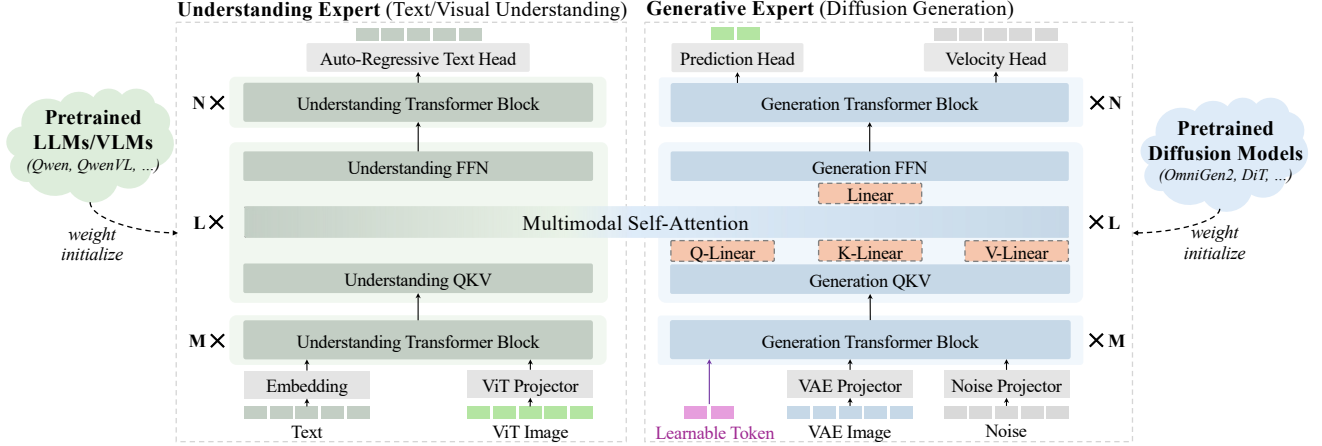


Figure 3. Overview of the proposed HBridge. We pair an arbitrary pretrained understanding expert with a pretrained generative expert and connect their mid-layers through self-attention. In practice, the understanding expert is typically a VLM, while the generative expert is often a DiT variant. QKV-Linear modules are applied to align their feature dimensions. Additionally, we introduce learnable semantic tokens that explicitly reconstruct visual semantic tokens of the target image, improving text alignment and enhancing generation quality.

first provide a brief overview of HBridge. Then, we will elaborate on the detailed mechanisms of each component.

3.1. Overview

Unlike prior symmetric MoT models [10, 28, 33, 45] that use identical Transformer branches with fully shared attention, HBridge introduces heterogeneous experts for understanding and generation and connects them only through a mid-layer semantic bridge. As shown in Fig. 3, the proposed HBridge introduces three key innovations: (1) Heterogeneous Experts, which allow the understanding and generation branches to adopt pretrained architectures better suited to their respective modalities. An understanding expert initialized from a large pretrained LLM or vision-language model, and a generative expert based on a diffusion-oriented Transformer; (2) Mid-layer Semantic Bridge, which enables selective information exchange, and only a subset of middle layers is connected via cross-attention bridges, forming an H-shaped topology that balances independence and interaction; (3) Semantic Reconstruction Tokens for explicit semantic grounding in generation. Together, these components enable effective semantic-visual alignment while preserving the distinct inductive biases and pretrained priors of both branches.

3.2. Heterogeneous Experts

Understanding Expert. The understanding branch is based on a large pretrained LLM/VLM [2, 48]. It encodes text and image inputs into high-level conceptual representations and performs token-level autoregressive semantic reasoning. This expert excels at contextual understanding and can serve as a semantic prior provider for visual generation. We explore two variants, named Qwen2.5-VL-7B [2] and Qwen2.5-0.5B [48]. To preserve the original powerful

reasoning capability of the pretrained model, we freeze the weights of the understanding expert and focus on improving the visual generative ability of the generative expert.

Generative Expert. The generative expert adopts a full-attention DiT-style diffusion model from OmniGen2 [58] pretrained for image synthesis. Since the understanding and generative experts differ significantly in internal configurations, such as embedding dimension, normalization type, and attention head number, direct attention sharing is infeasible. To resolve this, we introduce a QKV-Linear alignment module that enables the two experts to communicate within a unified latent space while preserving their own pretrained structures. Specifically, given QKV features $U_l^q \in \mathbb{R}^{L_u \times d_u^q}$, $U_l^k \in \mathbb{R}^{L_u \times d_u^k}$, $U_l^v \in \mathbb{R}^{L_u \times d_u^v}$ from the l -th layer of the understanding expert and $G_l^q \in \mathbb{R}^{L_g \times d_g^q}$, $G_l^k \in \mathbb{R}^{L_g \times d_g^k}$, $G_l^v \in \mathbb{R}^{L_g \times d_g^v}$ from the l -th generative expert, where L_u and L_g are token length, d_u and d_g mean the feature dimensions, we project the generation QKV features into a shared semantic space of understanding dimension:

$$Q_l = W_l^q G_l^q, \quad K_l = W_l^k G_l^k, \quad V_l = W_l^v G_l^v \quad (1)$$

where $W_l^q \in \mathbb{R}^{d_g^q \times d_u^q}$, $W_l^k \in \mathbb{R}^{d_g^k \times d_u^k}$, $W_l^v \in \mathbb{R}^{d_g^v \times d_u^v}$. Cross-modal attention is then performed on the resulting features. After semantic information interaction, the outputs of the generation branch are projected back to the original space of the diffusion expert through a linear operation. The diffusion model in [58] contains 32 Transformer layers, higher than the 28 layers in Qwen2.5-VL-7B and the 24 layers in Qwen2.5-0.5B. We move the redundant layers of the generative expert into the Noise Projector to ensure multimodal interaction with the understanding expert. To train the entire generative expert, we adopt the flow matching objective [30] following previous practices [10, 28, 29, 45].

3.3. Mid-layer Semantic Bridge

Full-layer attention sharing, as used in previous unified models [10, 28, 45], may ignore intrinsic task asymmetry: understanding tasks require high-level semantic reasoning, while generation needs to capture low-level details. The shared attention in the early input layer and the late output layer may hinder the learning of the modality-specific representations related to each task. In addition, we find that the dense connected MoT architecture may result in potential overfitting to shallow textual semantics, bypassing high-level reasoning semantics. Since many generative tasks, such as generating objects, can achieve promising results simply by utilizing shallow lexical or entity features from the frozen understanding expert during the training process, which causes the generative model to neglect extracting high-level reasoning cues. To alleviate this, our HBridge introduces a mid-layer semantic bridge, which connects the two experts only within a selected range of L intermediate layers, skipping the first M and last N connection layers.

3.4. Semantic Reconstruction Tokens

Generation tasks often require explicit semantic understanding, such as object relations, layout perception, and compositional reasoning. To enhance this, we introduce Semantic Reconstruction Tokens (SRT) on the input side of the generative expert. A small set of learnable tokens is appended to the generation input sequence. In the experiment, we apply 16 learnable tokens. During training, these tokens are supervised to reconstruct ViT-level semantic features of the ground-truth target image with cosine distance:

$$\mathcal{L}_{SRT} = \text{Distance}_{\cosine}(\text{Proj}(\text{Token}_{SRT}^{out}), F_{ViT}) \quad (2)$$

$$\mathcal{L} = \mathcal{L}_{Flowmatching} + \mathcal{L}_{SRT} \quad (3)$$

where F_{ViT} denotes adaptively pooled features extracted by a frozen pretrained ViT encoder of Qwen2.5-VL-7B to match the shape of the learnable tokens, and $\text{Proj}(\cdot)$ is a lightweight projection head. This auxiliary loss is jointly optimized together with the flow matching denoising loss. This semantic reconstruction objective explicitly injects semantic supervision into the generative process, encouraging the model to internalize relational semantics.

4. Experiments

In this section, we present a comprehensive quantitative and qualitative evaluation on various tasks to verify the effectiveness and superiority of the proposed HBridge.

4.1. Experimental Setup

Implementation Details. Our method involves two pre-trained experts. The understanding expert is initialized from a pretrained LLM/VLM backbone. Two variants, *i.e.*

Qwen2.5-0.5B [48] and Qwen2.5-VL-7B [2], are adopted to verify the effectiveness and implantability of the proposed method. The generative expert employs a 4B DiT from OminGen2 [58], and we add some linear layers to map the original dimension of the input tokens to match the attention operation, resulting in about 4B parameters. We skip the first of 6 layers and the last 6 layers, and only the middle layers are bridged for multimodal attention exchange. We use AdamW optimizer with a consistent learning rate of $1e-4$ to optimize our model. The total training step is about 200k. All experiments are implemented in PyTorch with mixed precision on 64 H100/A100/A800 GPUs. Some experiments are trained on 16 GPUs, and we use the gradient accumulation strategy to approximate training on 64 GPUs. We collected approximately 400M images from external open-source datasets [7, 29, 58, 75] and internal databases to train the model.

Benchmarks and Metrics. For multimodal understanding, since the understanding expert is frozen during the training process, the capability is preserved. For visual generation, we evaluate HBridge on a diverse set of generative benchmarks that cover both image generation (*i.e.*, DPG-Bench [16] and GenEval [14]) and image editing (*i.e.*, ImgEdit-Bench [66]). The benchmark metrics include counting, relationships, semantic understanding, *etc.*, and the final average result is used to indicate the final performance. Unless otherwise specified, our default setting is to use a **7B+4B** setting (understanding expert: Qwen2.5-VL-7B [2], generative expert: 4B DiT model [58]) for comparison with existing state-of-the-art methods. To effectively verify the effectiveness of our method and save computing resources, we adopt a **0.5B+4B** setting (understanding expert: Qwen2.5-0.5B [48], generative expert: 4B DiT model [58]) for ablation studies.

4.2. Comparison with State-of-the-Arts

Multimodal Visual Understanding. Our method leverages the Qwen2.5-VL-7B model as the understanding expert under the 7B+4B setting, which is frozen during training to fully preserve its native multimodal understanding capabilities. Like previous methods [29], our model can achieve solid scores of 83.5 on MMBench [32], 58.6 on MMMU [69], and 67.1 on MM-Vet [68], demonstrating strong visual-language reasoning performance.

Evaluation on Text-to-Image Generation. We evaluate the performance of text-to-image generation on two widely-used benchmarks, named DPG-Bench [16] and GenEval [14]. The results are shown in Tab. 1 and Tab. 2. From the results, we can observe that our method achieves state-of-the-art results over competitive counterparts, such as UniWorld-V1, OmniGen2, and BAGEL. Despite with only 4B generation parameters, HBridge can outperform BAGEL (7B+7B) and UniWorld-V1 (7B+12B), demon-

Table 1. Quantitative comparison of text-to-image generation results with state-of-the-art methods on the DPG-Bench benchmark [16].

Method	# Params.	Global ($\uparrow$)	Entity ($\uparrow$)	Attribute ($\uparrow$)	Relation ($\uparrow$)	Other ($\uparrow$)	Overall ($\uparrow$)
<i>Diffusion models / Autoregressive models</i>							
LUMINA-Next [78]	2B	82.82	88.65	86.44	80.53	81.82	74.63
SDXL [39]	2.6B	83.27	82.43	80.91	86.76	80.41	74.65
PlayGroundv2.5 [24]	-	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-DiT [27]	1.5B	84.59	80.59	88.01	74.36	86.41	78.87
PixArt- α [6]	0.6B	74.97	-	-	82.57	-	71.11
DALLE-3 [3]	-	90.97	89.61	88.39	90.58	89.83	83.50
SD3-medium [1]	2B	87.90	91.01	88.83	80.70	88.68	84.08
FLUX.1-dev [22]	12B	82.1	89.5	88.7	91.1	89.4	84.0
OmniGen [62]	3.8B	87.90	88.97	88.47	87.95	83.56	81.16
Infinity [15]	2B	93.11	-	-	90.76	-	83.46
SimpleAR [53]	1.5B	87.97	-	-	88.33	-	81.97
<i>Unified understanding and generation</i>							
Show-o [63]	1.3B	79.33	75.44	78.02	84.45	60.80	67.27
EMU3 [54]	8.5B	85.21	86.68	86.84	90.22	83.15	80.60
TokenFlow-XL [41]	14B	78.72	79.22	81.29	85.22	71.20	73.38
Janus [56]	1.5B	82.33	87.38	87.70	85.46	86.41	79.68
Janus Pro [8]	7B	86.90	88.90	89.40	89.32	89.48	84.19
BLIP3-o 4B [7]	3B + 1.4B	-	-	-	-	-	79.36
BLIP3-o 8B [7]	7B + 1.4B	-	-	-	-	-	81.60
UniWorld-V1 [29]	7B + 12B	83.64	88.39	88.44	89.27	87.22	81.38
OmniGen2 [58]	3B + 4B	88.81	88.83	90.18	89.37	90.27	83.57
BAGEL [10]	7B + 7B	88.94	90.37	91.29	90.82	88.67	85.07
HBridge	7B + 4B	91.78	91.82	90.23	90.06	88.42	85.23

strating the superiority and effectiveness of HBridge. It’s worth noting that compared to BAGEL trained with $\sim 2.5T$ T2I tokens, the proposed HBridge requires only $\sim 200B$ T2I tokens to achieve even better results.

Evaluation on Image Editing. As shown in Tab. 3, we evaluate the image editing performance of our method on ImgEdit-Bench [66]. The results indicate that HBridge achieves the best overall performance over all competitors, such as BAGEL, OmniGen2 and UniWorld-V1.

4.3. Ablation Study

We conduct ablation studies to validate the rationality of each component. Please refer to *appendix* for more results.

Effect of Pretrained Diffusion Expert. As displayed in Fig. 4, HBridge training with merely 40k steps can produce high-fidelity images. While replacing the pretrained diffusion backbone with a VLM initialized DiT leads to a sharp drop in generation quality even with more training steps. This confirms that our asymmetric heterogeneous expert design effectively leverages the pretrained diffusion priors while allowing flexible multimodal integration through QKV-Linear alignment. In addition, we can also observe that larger understanding experts (7B vs. 0.5B) can achieve better results with high visual quality.

Effect of Mid-layer Semantic Bridge. We conduct an ablation study on different values of M and N to verify the effectiveness of the mid-layer bridge. The quantitative



Figure 4. Ablation study on different initialization manners. We treat the understanding and generative experts initialized with Qwen2.5-VL-7B as the baseline, and our HBridge utilizes the pretrained diffusion initialization from [58].

comparisons are shown in Tab. 4, and $M=N=6$ achieves the best results on both DPG-Bench and GenEval. In Fig. 5, we display some visualization cases and can observe that when $M=N=10$, the semantics of some objects may be overlooked, such as “mushroom” and “sunflower”. We attribute this to the fact that there are too few semantic injection layers for visual-text alignment, making it difficult

Table 2. Text-to-image results on GenEval benchmark [14]. “*” means reproducing the result using the official open-source checkpoints.

Method	Single object (↑)	Two object (↑)	Counting (↑)	Colors (↑)	Position (↑)	Color attri. (↑)	Overall (↑)
<i>Diffusion models / Autoregressive models</i>							
LUMINA-Next [78]	0.92	0.46	0.48	0.70	0.09	0.13	0.46
SD3-medium [1]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
FLUX.1-dev [22]	0.99	0.81	0.79	0.74	0.20	0.47	0.67
NOVA [11]	0.99	0.91	0.62	0.85	0.33	0.56	0.71
OmniGen [62]	0.98	0.84	0.66	0.74	0.40	0.43	0.68
Infinity [15]	-	0.85	-	-	0.49	0.57	0.73
SimpleAR [53]	-	0.90	-	-	0.28	0.45	0.63
<i>Unified understanding and generation</i>							
Chameleon [46]	-	-	-	-	-	-	0.39
TokenFlow-XL [41]	0.95	0.60	0.41	0.81	0.16	0.24	0.55
Janus [56]	0.97	0.68	0.30	0.84	0.46	0.42	0.61
ILLUME [51]	0.99	0.86	0.45	0.71	0.39	0.28	0.61
Transfusion [76]	-	-	-	-	-	-	0.63
Janus Pro [8]	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Show-o [63]	0.98	0.80	0.66	0.84	0.31	0.50	0.68
UniWorld-V1 [29]	0.99	0.93	0.79	0.89	0.49	0.70	0.80
OmniGen2 [58]	1	0.95	0.64	0.88	0.55	0.76	0.80
Nexus-Gen(7B+12B) [70]	0.97	0.93	0.64	0.88	0.83	0.62	0.81
BAGEL* [10]	1	0.94	0.79	0.88	0.55	0.66	0.80
HBridge	0.97	0.94	0.73	0.93	0.63	0.78	0.83
<i>Unified understanding and generation with LLM re-writer</i>							
Emu3-Gen [54]	0.99	0.81	0.42	0.80	0.49	0.45	0.66
MetaQuery-XL [37]	-	-	-	-	-	-	0.80
BLIP3-o 4B [7]	-	-	-	-	-	-	0.81
BLIP3-o 8B [7]	-	-	-	-	-	-	0.84
UniWorld-V1 [29]	0.98	0.93	0.81	0.89	0.74	0.71	0.84
OmniGen2 [58]	0.99	0.96	0.74	0.98	0.71	0.75	0.86
BAGEL* [10]	0.98	0.95	0.81	0.92	0.73	0.74	0.86
Hyper-BAGEL [33]	0.99	0.94	0.86	0.94	0.72	0.74	0.86
HBridge	1	0.96	0.80	0.94	0.77	0.78	0.87

Table 3. Quantitative results of image editing on ImgEdit-Bench [66]. “Overall” indicates the average score across all editing sub-tasks.

Method	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall (↑)
MagicBrush [71]	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
Instruct-P2P [4]	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.2	1.46	1.88
AnyEdit [67]	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
UltraEdit [75]	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.7
Step1X-Edit [31]	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
ICEdit [74]	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
UniWorld-V1 [29]	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
OmniGen2 [58]	3.57	3.06	1.77	3.74	3.2	3.57	4.81	2.52	4.68	3.44
BAGEL [10]	3.56	3.31	1.7	3.3	2.62	3.24	4.49	2.38	4.17	3.20
HBridge	3.91	3.23	2.13	3.79	3.21	3.50	4.31	3.44	4.32	3.54

to learn reliable prompt following. When $M=N=0$, the results remain visually realistic but may fail to match certain attributes, relations, and numerical terms, such as “taller”, “three”, etc. In Fig. 6, we can notice that when $M=N=0$, there is an overfitting phenomenon here, where the shallow semantic layer dominates the generation process, which makes it difficult to understand some complex prompts, since complex semantic understanding is generally in the middle and deep layers of a large understanding model [13,

21]. During the training process, many generative tasks, such as generating objects, can achieve promising results simply by utilizing shallow lexical or entity features from the frozen understanding expert, which causes the generative model to neglect extracting high-level reasoning cues. Among these, $M=N=6$ achieves a good balance.

Effect of Semantic Reconstruction Tokens. As illustrated in Tab. 5, the proposed semantic reconstruction tokens

Table 4. Ablation study on the number of skipped layers under the **0.5B+4B** setting. LLM re-writer is not used on GenEval.

Setting	DPG-Bench	GenEval
M=0, N=0	78.19	0.61
M=2, N=2	78.63	0.62
M=4, N=4	79.20	0.66
M=0, N=6	79.36	0.63
M=6, N=6	80.03	0.66
M=6, N=0	79.26	0.64
M=8, N=8	73.51	0.54
M=10, N=10	69.05	0.35

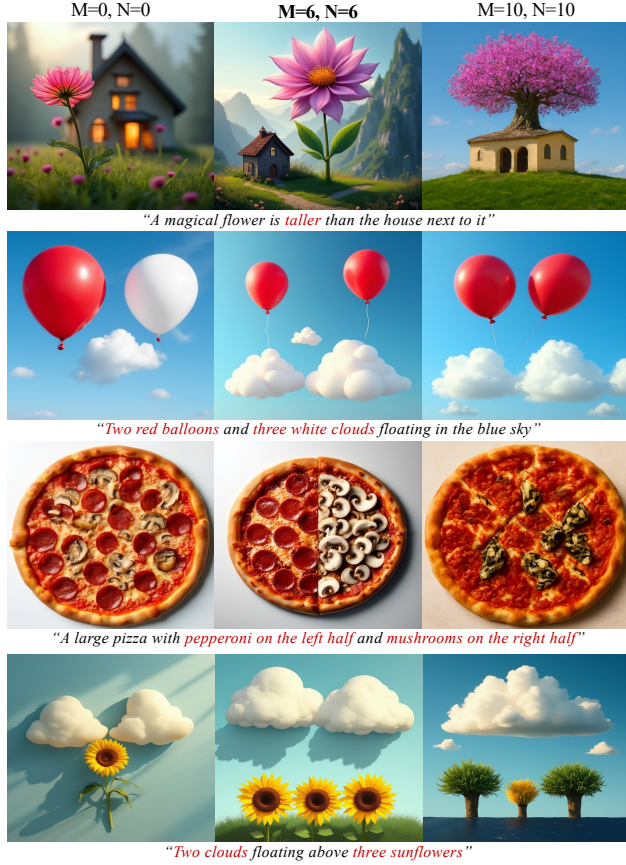


Figure 5. Ablation study on varying the number of skipped layers.

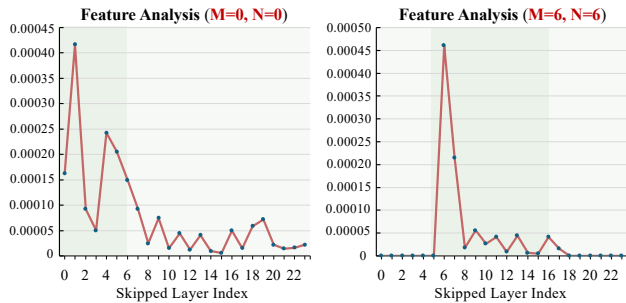


Figure 6. Analysis of varying the skipped layer on DPG-Bench.

help to improve the performance. We attribute to that the learnable tokens explicitly inject semantic supervision

Table 5. Ablation study on semantic reconstruction tokens.

Setting	DPG-Bench	GenEval
w/o semantic tokens	79.57	0.65
w/ semantic tokens (Ours)	80.03	0.66

Table 6. Ablation study on different cross-expert fusion manners.

Setting	DPG-Bench	GenEval
Shallow fusion	74.53	0.58
Deep fusion (Ours)	80.03	0.66

Table 7. Generalization experiments based on BAGEL (7B+7B). “MSB” and “SRT” are abbreviations for Mid-layer Semantic Bridge and Semantic Reconstruction Tokens, respectively. “†” refers to the finetuned method using our collected data.

Method	DPG-Bench	GenEval	ImgEdit-Bench
BAGEL [10]	85.07	0.80	3.20
BAGEL [†]	85.20	0.82	3.25
BAGEL + MSB	85.41	0.84	3.34
BAGEL + MSB + SRT	85.55	0.84	3.38

by reconstructing ViT features, enforcing a stronger link between textual semantics and visual generation.

Shallow Fusion vs. Deep Fusion. Many recent advanced methods [29, 57] use the output embeddings of the last layer in LLMs/VLMs to guide subsequent diffusion model. We further compare this shallow fusion strategy with our deep fusion manner. The comparison is exhibited in Tab. 6. The deep fusion manner achieves better performance, and we believe the reason is that deep fusion can utilize rich, multi-layer semantic information from the understanding expert, while shallow fusion can only utilize the last layer.

Pluggability. As shown in Tab. 7, we insert the proposed mid-layer semantic bridge and semantic reconstruction tokens into the open-source pretrained BAGEL and find that performance can be further improved, validating the insertability and generalization of the proposed components.

5. Conclusion and Limitations

In this work, we presented HBridge, a unified H-shaped framework that integrates multimodal understanding and generation through an asymmetric MoT architecture. By introducing heterogeneous experts, a mid-layer semantic bridge, and semantic reconstruction tokens, HBridge achieves superior performance with low training cost.

Limitations. Despite its effectiveness, HBridge still faces several limitations: 1) to reduce experimental costs, we usually assume that $M=N$ and that both are even numbers in our experiments, and more detailed settings for different combinations of M and N are not explored; 2) the mid-layer bridge currently uses fixed coupling layers. Dynamic layer selection based on semantic salience may further improve performance and flexibility in the future.

References

- [1] Stability AI. Sd3-medium. <https://stability.ai/news/stable-diffusion-3-medium>, 2024. 3, 6, 7
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 4, 5
- [3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023. 3, 6
- [4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 3, 7
- [5] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xincheng Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025. 2, 3
- [6] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 3, 6
- [7] Jiahai 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. 2, 3, 5, 6, 7
- [8] 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. 2, 6, 7
- [9] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *CVPR*, pages 12501–12511, 2025. 3
- [10] 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. 2, 3, 4, 5, 6, 7, 8, 13
- [11] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024. 7
- [12] 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 *ICML*, 2024. 2, 3
- [13] Siqi Fan, Xin Jiang, Xiang Li, Xuying Meng, Peng Han, Shuo Shang, Aixin Sun, Yequan Wang, and Zhongyuan Wang. Not all layers of llms are necessary during inference. *arXiv preprint arXiv:2403.02181*, 2024. 7
- [14] Dhruva Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, pages 52132–52152, 2023. 5, 7
- [15] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *CVPR*, pages 15733–15744, 2025. 6, 7
- [16] 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. 5, 6
- [17] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: creative and controllable image synthesis with composable conditions. In *ICML*, pages 13753–13773, 2023. 3
- [18] 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. 2
- [19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 3
- [20] Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In *CVPRW*, pages 3600–3610, 2025. 3
- [21] Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. How large language models encode context knowledge? a layer-wise probing study. *arXiv preprint arXiv:2402.16061*, 2024. 7
- [22] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 3, 6, 7
- [23] Ang Li, Ben Liu, Binbin Hu, Bing Li, Bingwei Zeng, Borui Ye, Caizhi Tang, Changxin Tian, Chao Huang, Chao Zhang, et al. Every activation boosted: Scaling general reasoner to 1 trillion open language foundation. *arXiv preprint arXiv:2510.22115*, 2025. 3
- [24] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 6
- [25] Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. Onecat: Decoder-only auto-regressive model for unified understanding and generation. *arXiv preprint arXiv:2509.03498*, 2025. 3
- [26] Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. Llava-o: Elastic masked diffusion models for unified multimodal understanding and generation. *arXiv preprint arXiv:2509.19244*, 2025. 2
- [27] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xincheng Deng, Yingfang Zhang, Xingchao Liu,

- Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 6
- [28] 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. 2, 3, 4, 5
- [29] 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. 3, 4, 5, 6, 7, 8, 13
- [30] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3, 4
- [31] 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. 3, 7
- [32] 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 *ECCV*, pages 216–233. Springer, 2024. 5
- [33] Yanzuo Lu, Xin Xia, Manlin Zhang, Huafeng Kuang, Jianbin Zheng, Yuxi Ren, and Xuefeng Xiao. Hyper-bagel: A unified acceleration framework for multimodal understanding and generation. *arXiv preprint arXiv:2509.18824*, 2025. 2, 4, 7
- [34] 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 *CVPR*, pages 7739–7751, 2025. 3
- [35] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487*, 2025. 3
- [36] Sicheng Mo, Thao Nguyen, Xun Huang, Siddharth Srivasan Iyer, Yijun Li, Yuchen Liu, Abhishek Tandon, Eli Shechtman, Krishna Kumar Singh, Yong Jae Lee, et al. X-fusion: Introducing new modality to frozen large language models. In *ICCV*, 2025. 2
- [37] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiahui Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 2, 3, 7
- [38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 3
- [39] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3, 6
- [40] Attilio A Poli and Mario C Cirillo. On the use of the normalized mean square error in evaluating dispersion model performance. *Atmospheric Environment. Part A. General Topics*, 27(15):2427–2434, 1993. 14
- [41] 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 *CVPR*, pages 2545–2555, 2025. 3, 6, 7
- [42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022.
- [43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3
- [44] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *CVPR*, pages 8871–8879, 2024. 3
- [45] 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. 2, 3, 4, 5
- [46] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2, 3, 7
- [47] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025. 3
- [48] Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024. 4, 5
- [49] Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In *ICCV*, pages 17001–17012, 2025. 2, 3
- [50] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2
- [51] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. In *ICCV*, pages 21612–21622, 2025. 7
- [52] Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, et al. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025. 3
- [53] Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing

- the frontier of autoregressive visual generation through pre-training, sft, and rl. *arXiv preprint arXiv:2504.11455*, 2025. 6, 7
- [54] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yuezhang Wang, Zhen Li, Qiyang Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 2, 3, 6, 7
- [55] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.
- [56] 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 *CVPR*, pages 12966–12977, 2025. 6, 7
- [57] 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. 8
- [58] Chenyuan Wu, Pengfei Zheng, Ruirao Yan, Shitao Xiao, Xin Luo, Yuezhang Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 2, 4, 5, 6, 7, 13
- [59] Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*, 2025.
- [60] 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.
- [61] 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. 2
- [62] Shitao Xiao, Yuezhang Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruirao Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *CVPR*, pages 13294–13304, 2025. 6, 7
- [63] 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. 2, 3, 6, 7
- [64] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 3
- [65] Junzhe Xu, Yuyang Yin, and Xi Chen. Tbac-uniimage: Unified understanding and generation by ladder-side diffusion tuning. *arXiv preprint arXiv:2508.08098*, 2025. 2
- [66] 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. 3, 5, 6, 7
- [67] 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 *CVPR*, pages 26125–26135, 2025. 7
- [68] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 5
- [69] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoyi 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 *CVPR*, pages 9556–9567, 2024. 5
- [70] Hong Zhang, Zhongjie Duan, Xingjun Wang, Yuze Zhao, Weiyi Lu, Zhipeng Di, Yixuan Xu, Yingda Chen, and Yu Zhang. Nexus-gen: A unified model for image understanding, generation, and editing. *arXiv preprint arXiv:2504.21356*, 2025. 2, 7
- [71] Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS*, 2024. 7
- [72] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 3
- [73] Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Jiakui Hu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025. 2
- [74] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025. 7
- [75] Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. In *NeurIPS*, pages 3058–3093, 2024. 3, 5, 7
- [76] 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. 2, 3, 7
- [77] Xianwei Zhuang, Yuxin Xie, Yufan Deng, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian Zou. Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. *arXiv preprint arXiv:2501.12327*, 2025. 2
- [78] Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. In *NeurIPS*, pages 131278–131315, 2024. 6, 7

HBridge: H-Shape Bridging of Heterogeneous Experts for Unified Multimodal Understanding and Generation

Supplementary Material



Figure 7. More qualitative results of text-to-image generation synthesized by the proposed HBridge.

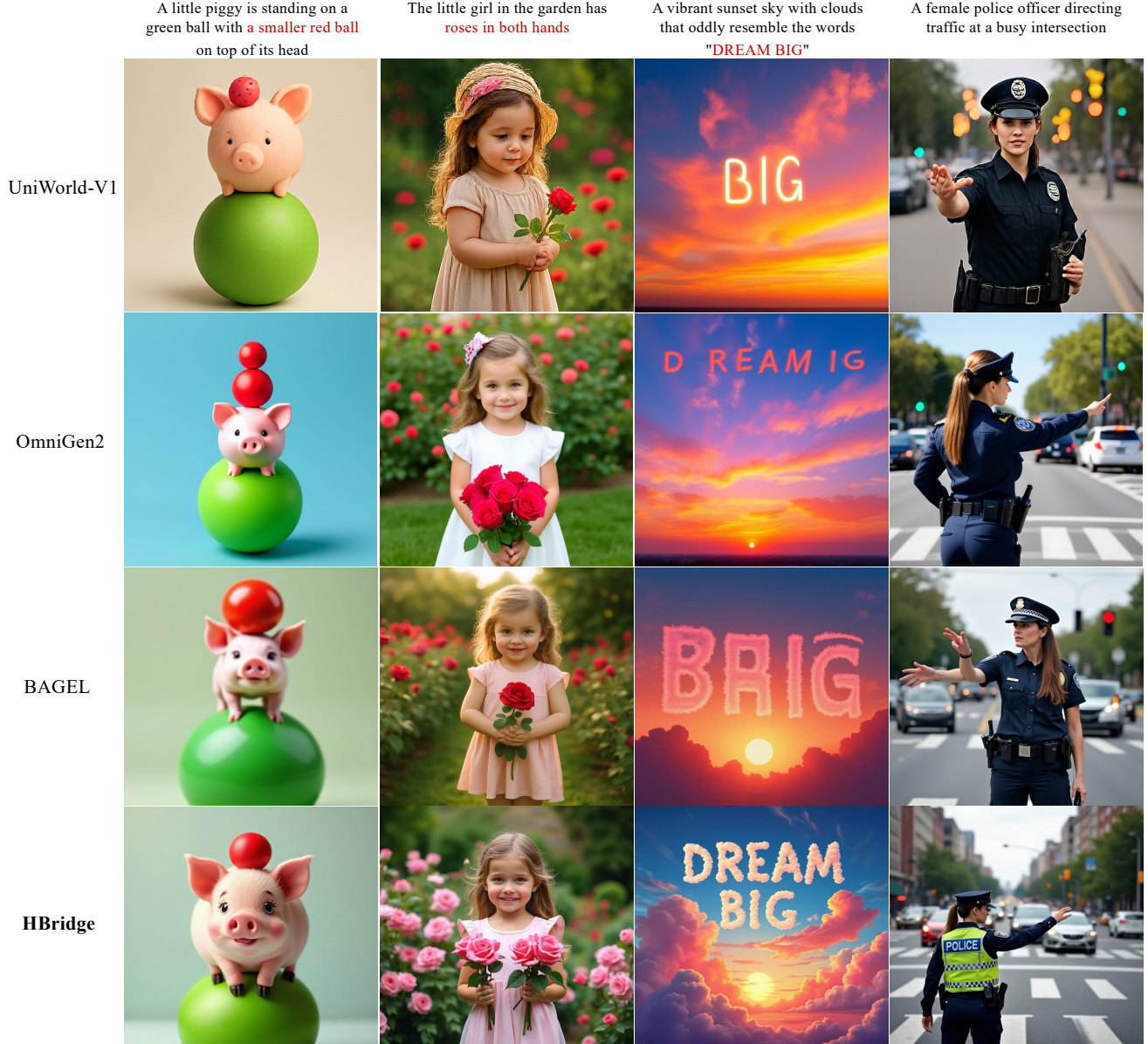


Figure 8. Comparison with the state-of-the-art methods such as UniWorld-V1, OmniGen2 and BAGEL on text-to-image generation task.

Due to the page limit of the main document, we place some supplementary results and details in the appendix.

6. More Qualitative Results

As exhibited in Fig. 7, we show more high-quality, photo-realistic text-to-image cases generated by HBridge with different resolution rates. These examples demonstrate excellent spatial layout, quantity control, and text rendering, validating powerful generative capabilities of HBridge.

In addition, we qualitatively compare our method with the state-of-the-art methods, including UniWorld-V1 [29], OmniGen2 [58], and BAGEL [10]. The text-to-image

results are shown in Fig. 8. Our method demonstrates better semantic coherence and visual quality. The editing results are displayed in Fig. 9, and HBridge can precisely understand and respond to the user’s intentions, resulting in reliable editing outcomes. We attribute this to the fact that our H-shape design with semantic tokens helps improve the semantic understanding capabilities of generative models.

7. Additional Ablation Results

Overfitting Phenomenon under 7B+4B Settings. To further verify that the fully layer-by-layer connected method may easily overfit the shallow features of the understanding



Figure 9. Comparison with state-of-the-art methods on image editing tasks, including object removal and replacement.

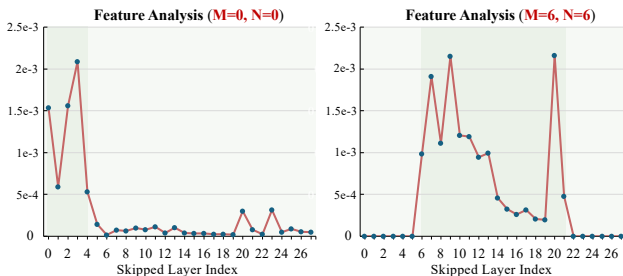


Figure 10. Analysis of varying the skipped layer under the **7B+4B** setting with 28 Transformer layers. We disconnect the multimodal self-attention layer by layer and analyze the differences in output features caused by disconnecting and reconnecting the multimodal self-attention layer. These differences are measured using the average normalized MSE [40] on DPG-Bench.

Table 8. Ablation study on the number of skipped layers under the **7B+4B** setting. LLM re-writer is not used on GenEval.

Setting	DPG-Bench	GenEval
M=0, N=0	83.21	0.80
M=6, N=6 (Ours)	85.23	0.83



Figure 11. Qualitative ablation study on the number of skipped layers under the **7B+4B** setting.

expert, we conduct additional experiments under the **7B+4B** configuration. From the result in Fig. 10, it can be seen that $M=N=0$ easily leads to overfitting of shallow features, while the setting of **HBridge** focuses primarily on the features of the intermediate semantic layers, resulting in better semantic coherence. We also show the quantitative results under the **7B+4B** setting in Tab. 8, we can find that the mid-layer bridge ($M=N=6$) performs better than the baseline counterpart ($M=N=0$). In addition, the qualitative results in Fig. 11 demonstrate that $M=N=0$ may ignore some high-level semantics in textual prompts. The conclusions are consistent with those under the **0.5B+4B** setting in the main document.

Effect of Semantic Reconstruction Tokens. To qualitatively analyze the efficacy of the proposed semantic recon-

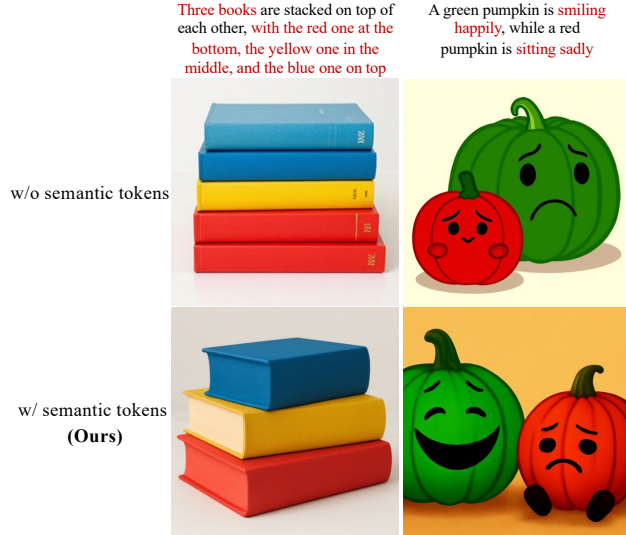


Figure 12. Qualitative ablation study on the effect of the proposed semantic reconstruction tokens.

Table 9. Ablation study on the number of learnable semantic tokens under the 0.5B+4B setting.

Setting	DPG-Bench	GenEval
4 tokens	79.82	0.65
16 tokens	80.03	0.66
36 tokens	79.83	0.66

struction tokens, we visualize some examples in Fig. 12. From the results, we can notice that incorporating semantic reconstruction tokens helps to enhance the ability to perceive position and attributes. As shown in Tab. 9, we further conduct an ablation study on the number of learnable semantic tokens and find that 16 tokens achieve excellent performance.