# SPAR: Semantic-Pixel Self-Alignment and Adaptive Routing for Unified Multimodal Models

Hongxiang Li[1*], Hongxu Chen[1*], Chenyang Zhu[1], Xiaoshuang Huang[2], Jiayin Cai[2], Xiaolong Jiang[2], Yao Hu[2], and Long Chen[1†]

[1] The Hong Kong University of Science and Technology
[2] Xiaohongshu Inc.
hlihg@connect.ust.hk, longchen@ust.hk
Project page: https://hkust-longgroup.github.io/SPAR/

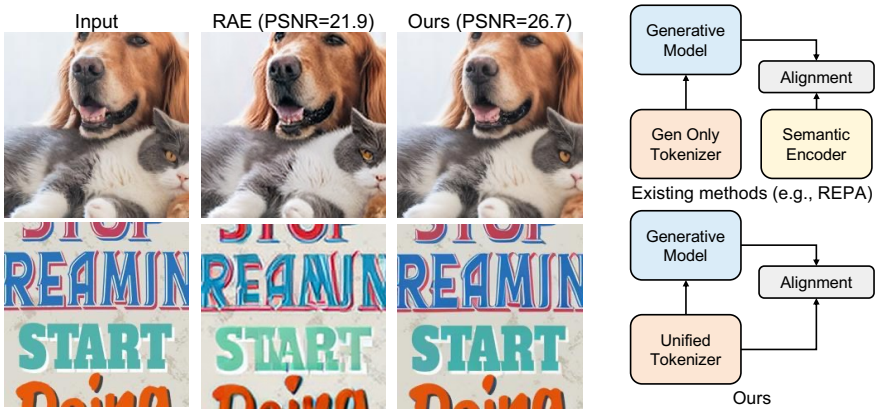
**Abstract.** Multimodal Large Language Models (MLLMs) have achieved remarkable success in visual understanding but remain constrained in visual generation due to the fundamental feature discrepancy between semantic perception and pixel-level reconstruction. Bridging this gap requires overcoming two core challenges: endowing semantic encoders with high-fidelity reconstruction capabilities, and effectively aligning generative models with semantic spaces without relying on external teachers. To this end, we propose a novel unified multimodal framework featuring **S**emantic-**P**ixel self-alignment and **A**daptive **R**outing (**SPAR**). First, to reconcile semantic perception with pixel-level reconstruction, we introduce an asymmetric dual-stream unified tokenizer. A lightweight semantic stream anchors discriminative features, while a Transformer-augmented pixel stream recovers fine-grained visual details into a unified compact latent space. Second, to eliminate external dependencies, we propose a self-aligned generation paradigm that natively leverages this optimized tokenizer as an internal alignment teacher for the diffusion model. Furthermore, to facilitate flexible multimodal interaction within this unified space, we introduce Dynamic Token Routing, which enables each token to adaptively aggregate multi-layer MLLM features based on its distinct semantic demands. Extensive experiments demonstrate that SPAR establishes the state-of-the-art for unified architectures, achieving exceptional generation and reconstruction quality while preserving foundational visual understanding capabilities.

**Keywords:** Unified Model · Visual Generation · Visual Tokenizer

## 1 Introduction

Multimodal Large Language Models (MLLMs) [1, 28, 53] have demonstrated exceptional performance across a wide range of perception tasks, such as image captioning, visual question answering, and multimodal reasoning, by aligning

* Equal contribution.
† Corresponding author.




(a) Image Reconstruction (b) Representation Alignment

**Fig. 1: (a) Image Reconstruction:** When modeling directly within the semantic representation space, existing methods suffer from lossy compression and struggle to preserve high-frequency details. In contrast, our method effectively recovers these crucial pixel-level details. **(b) Representation Alignment Paradigm:** Unlike existing approaches that rely on external semantic encoders to guide the generative model, our framework natively employs the unified tokenizer itself as an alignment teacher.

powerful visual semantic encoders with Large Language Models (LLMs) [51,68]. While these models have become pervasive paradigms in the understanding domain, they have yet to dominate the field of visual generation. Meanwhile, current state-of-the-art visual generation systems [6, 20, 21, 25, 55–57, 59] still rely primarily on low-level, compact latent spaces constructed by Variational Auto-Encoders (VAEs) [18], upon which diffusion processes [12, 39] or autoregressive modeling [48, 72] are performed. This leads to a discrepancy in feature patterns between perception and generation models.

Recent studies [69, 73, 82] have explored incorporating structured semantic representations (*e.g.*, DINOv2 [35]) into generative models. By either aligning diffusion processes with these rich semantic priors or developing generative models directly upon them, these approaches have yielded significant gains in generation quality and convergence speed. The growing efficacy of these semantic spaces in generation reveals a promising path toward unifying visual perception and generation. This raises **a natural and critical question**: *Rather than relying on latent spaces, can we seamlessly empower powerful MLLMs with image generation capabilities by shifting the generative modeling space directly to the semantic representation space?*

However, realizing this paradigm migration is far from straightforward, as an inherent contradiction between semantic perception and pixel reconstruction needs to be overcome. Although recent approaches [69, 82] have successfully implemented diffusion models directly within the semantic representation space, they suffer from severe limitations in image reconstruction quality. As illustrated in Figure 1(a), RAE [82] struggles to preserve high-frequency details,

often yielding blurry artifacts and distorted structures. This occurs because semantic encoders optimized through representation learning have feature spaces that are highly converged toward discriminative tasks. This process is essentially a lossy compression of the input image, which tends to discard texture and structural details that are irrelevant to perception. But these details may be crucial for pixel reconstruction. Consequently, this inherently weak reconstruction capability severely hinders the model's synthesis quality when scaled to text-to-image generation and complex instruction-based image editing [78]. Furthermore, completely fine-tuning the semantic encoder to recover these lost pixel-level details would inevitably trigger catastrophic forgetting, destroying its original understanding abilities. Therefore, *how to endow the semantic encoder with high-quality pixel reconstruction capability without compromising its original understanding ability* remains the core bottleneck for migrating the generative modeling space to the semantic representation space.

Moreover, another critical challenge arises when effectively aligning representations to the generative model during unified model training. As shown in Figure 1(b), existing representation-aligned generation methods [69,73] typically bridge this gap by enforcing the diffusion model to align with external visual encoders [35] While this forced external alignment provides structured guidance, recent studies [3] suggest that as data scale grows, relying on an isolated external teacher becomes sub-optimal. It risks manifold mismatches between the multimodal understanding and visual generation spaces. Consequently, *how to natively align the generative model directly with the unified semantic space, and eliminate the dependency on external representation learners* is a key challenge for effectively training unified understanding and generation models.

To address the above challenges, we propose Semantic-Pixel self-alignment and Adaptive Routing (**SPAR**) for unified multimodal models. To overcome the inherent contradiction between semantic perception and pixel reconstruction, we introduce an asymmetric dual-stream self-alignment architecture into the unified tokenizer. The lightweight semantic stream acts as an anchor to preserve the encoder's original discriminative features, preventing catastrophic forgetting. Concurrently, the Transformer-augmented pixel stream is dedicated to mapping the high-dimensional semantic space into a compact latent space, effectively recovering the high-frequency spatial textures discarded by the encoder's lossy compression. The two streams are fused in a compact latent space, explicitly decoupling semantic preservation from pixel reconstruction. Notably, the dual-stream-optimized tokenizer itself can serve as the representation alignment teacher for the diffusion model, seamlessly transferring the accumulated semantic and pixel knowledge to the generation stage without relying on any disjointed external feature model. Furthermore, we introduce Dynamic Token Routing (DTR) to facilitate flexible multimodal interaction within this unified space. It empowers each token to adaptively aggregate multi-layer MLLM features based on its distinct semantic role, dynamically satisfying the differentiated hierarchical feature demands of varying tokens.

In summary, our contributions are threefold:

- We propose a semantic-pixel dual-stream self-alignment architecture that explicitly decouples semantic preservation and pixel reconstruction through an asymmetric design, achieving high-fidelity reconstruction without degrading the encoder's understanding capability.
- We introduce dynamic token routing, which enables each token to adaptively determine its own multi-layer fusion weights from the MLLM to optimally harness multimodal representations for generative guidance.
- We present a self-alignment generation paradigm that leverages the unified tokenizer as an alignment teacher, removing the dependency on external feature models while achieving competitive results across multiple benchmarks.

## 2 Related Work

**Unified Visual Tokenizer.** Early approaches [52, 83] relied on reconstruction-focused generative models, which often lack the high-level semantics required for understanding tasks. To address this, recent works have explored unified tokenizers from various perspectives. Some attempt to discretize semantic features directly [32, 37, 63], but suffer from information loss inherent in quantization. Another recent methods [40, 44, 46, 76, 82] attempt to endow pre-trained semantic encoders with pixel reconstruction capabilities via self-distillation and continuous feature alignment. However, forcing a single representation stream to simultaneously accommodate abstract semantics and fine-grained spatial details inevitably leads to capacity bottlenecks and implicit performance trade-offs. To resolve this tension, we explicitly decouple these competing objectives through an unified tokenizer, coordinating visual understanding and generation.

**Representation-Guided Generation.** Pre-trained visual representations are increasingly integrated into diffusion models to enhance semantic and structural fidelity. Recent methods typically achieve this by either explicitly aligning the generative latent space with external discriminative features [35, 73], or directly performing diffusion modeling within the semantic space [82]. However, these approaches face inherent limitations. As highly abstract semantic spaces naturally discard pixel details, such models often suffer from degraded reconstruction quality and off-manifold artifacts [78]. Furthermore, extracting guidance features from external networks forces the generative process to rely on a representation space that is entirely isolated from the model's native visual encoder. To address this, we introduce a self-aligned generation paradigm. By natively utilizing our explicitly decoupled tokenizer as an internal alignment objective.

**Unified Multimodal Understanding and Generation.** The integration of visual understanding and generation within a single LLM has garnered significant attention [**?**, 8, 11, 24, 81, 84]. Early unified models [47, 70] typically formulated both tasks as next-token prediction within a purely autoregressive framework. To leverage the superior synthesis quality of diffusion models, recent approaches propose attaching diffusion modules directly to the LLM backbone [45, 66]. To bridge the modality gap, these architectures employ various connection mechanisms. For instance, some methods utilize learnable Q-Formers or linear projectors to align representations [7, 61], while others adopt Mixture-of-Experts

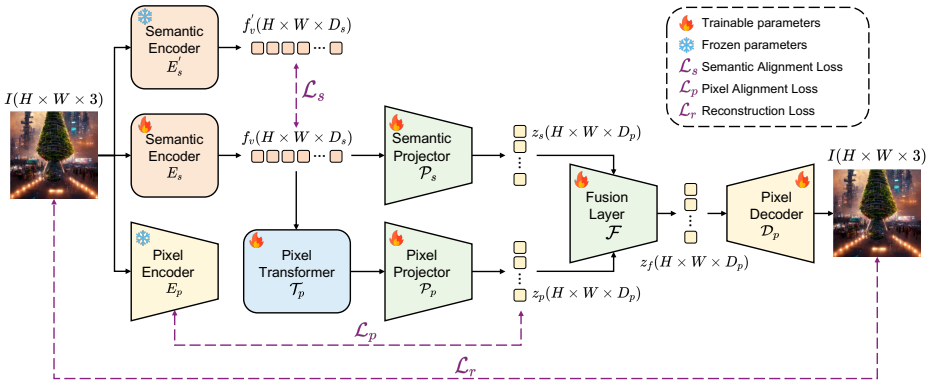



**Fig. 2: Architecture of the semantic-pixel self-aligned unified tokenizer.** Our tokenizer design explicitly decouples semantic preservation from pixel reconstruction. A lightweight semantic stream maps features into a compact latent space, strictly anchored by the frozen encoder to prevent catastrophic forgetting ($\mathcal{L}_s$). Concurrently, a Transformer-augmented pixel stream bridges the dimensional gap by aligning with the native pixel latent space ($\mathcal{L}_p$) to recover high-frequency spatial details. Both streams are fused and decoded to reconstruct the image ($\mathcal{L}_r$), endowing the model with high-fidelity generation capabilities without compromising its discriminative understanding.

(MoE) or specialized visual queries to manage task routing [11, 26, 46]. Despite diverse connector designs, a commonality among these models is their reliance on static, token-agnostic feature extraction (*e.g.*, using only the final layer or fixed fusion weights). This rigid paradigm fails to fully exploit the rich hierarchical representations within MLLMs. To address this, SPAR introduces dynamic token routing to adaptively aggregate multi-layer features for optimal generative guidance.

## 3 Method

### 3.1 Semantic-Pixel Self-Aligned Unified Tokenizer

The unified tokenizer aims to construct a visual encoding space that preserves both discriminative semantic information and pixel-level reconstruction capability. Consequently, it can serve as the visual encoder of the MLLM for understanding tasks while simultaneously providing a compact latent space for diffusion modeling. Training pixel decoders directly on the frozen semantic encoder only yields blurry reconstruction results, while unlocking the encoder for reconstruction training triggers catastrophic forgetting.

Semantic information has already been sufficiently encoded in the output features of the pre-trained encoder, and applying excessive transformations would instead perturb the original feature distribution. In contrast, the pixel-level details discarded by the encoder during representation learning require additional computational capacity for recovery. Based on this asymmetry, we explicitly

decouple the two competing objectives of semantic preservation and pixel reconstruction as shown in Fig. 2. For the input image $I \in \mathbb{R}^{H\times W\times 3}$, we construct two different streams on top of the encoder output feature $f_v = E_s(I) \in \mathbb{R}^{H^{'}\times W^{'}\times D_s}$, where $E_s$ is semantic encoder and $D_s$ is the feature dimension.

**Semantic Stream.** The semantic stream employs a lightweight projector $\mathcal{P}_s$ consisting of multiple residual blocks and an MLP projection layer, which maps the encoder features to the compact latent space with minimal transformation: $\mathbf{z}_\mathrm{s} = \mathcal{P}_s(f_v) \in \mathbb{R}^{H^{'}\times W^{'}\times D_p}$, where $D_p \ll D_s$ is the latent space dimension. The residual connections ensures the output of the semantic stream to preserve the original discriminative structure of $f_v$ as much as possible, serving as an anchor for the encoder's semantic features in the latent space.

**Pixel Stream.** A fundamental gap exists between the feature space of the semantic encoder and the compact latent space operated on by the pixel decoder. The semantic space is high-dimensional ($D_s$) and optimized for discriminative objectives, whereas the pixel latent space is low-dimensional ($D_p \ll D_s$) and tailored for reconstruction. For the semantic encoder's output to be correctly decoded by the pixel decoder, an effective mapping from the semantic feature space to the pixel latent space must first be established. The pixel stream is designed as a learnable bridge for this purpose. It transfers high-dimensional semantic features into the compact pixel latent space, thus allowing the fine-grained details discarded by the encoder during representation learning to be recovered within the pixel decoder's native space. Specifically, the pixel stream is equipped with an independent Transformer encoder $\mathcal{T}_p$, which models long-range dependencies along the spatial dimension through global self-attention and transforms the semantic features into intermediate representations suitable for pixel reconstruction:

$$f_p = \mathcal{T}_p(f_v) \in \mathbb{R}^{H^{'}\times W^{'}\times D_s}.$$

The Transformer output $f_p$ is then mapped to a compact latent space compatible with the pixel decoder through residual blocks and an MLP projection layer:

$$\mathbf{z}_\mathrm{p} = \mathcal{P}_p(f_p) \in \mathbb{R}^{H^{'}\times W^{'}\times D_p}.$$

**Fusion & Decoding**. The compact latent variables produced by the two streams are merged into a unified representation in the fusion layer $\mathcal{F}$. The semantic latent $\mathbf{z}_\mathrm{s}$ and the pixel latent $\mathbf{z}_\mathrm{p}$ are concatenated along the channel dimension and then mapped back to the latent space dimension by the fusion layer:

$$\mathbf{z}_f = \mathcal{F}\big([\mathbf{z}_\mathrm{s};\, \mathbf{z}_\mathrm{p}]\big) \in \mathbb{R}^{H^{'}\times W^{'}\times D_p},$$

where $[\cdot;\cdot]$ denotes channel concatenation. The fused $\mathbf{z}_f$ is fed into the pre-trained pixel decoder $\mathcal{D}_p$ for image reconstruction. This design allows semantic and pixel information to evolve along their respective independent optimization paths before converging in the compact latent space.

### 3.2 Progressive Three-Stage Training for Unified Tokenizer

The training of the unified tokenizer follows a progressive three-stage strategy that gradually releases model capacity, with targeted alignment losses introduced at each stage to guide the optimization direction.

**Stage I: Dual-Stream Initialization.** The vision encoder and pixel decoder are frozen, and only the dual-stream modules are trained. The core objective of this stage is to enable the pixel stream to learn the transfer mapping from the semantic feature space to the pixel latent space. However, the semantic feature space and the pixel decoder's latent space differ fundamentally in dimensionality, distribution, and optimization objectives, making it difficult for the pixel stream to discover the correct mapping direction without explicit supervision. To this end, we introduce the **pixel alignment loss** that uses the output $\mathbf{z}'_p = E'_p(I)$ of the frozen pixel encoder $E_p$ as an explicit optimization anchor and aligns the pixel stream's latent variables with it:

$$\mathcal{L}_p = \|\mathbf{z}_p - \mathbf{z}'_p\|_2^2.$$

This loss first guides the pixel stream to transfer the semantic feature space onto the latent manifold of the pixel encoder, aligning its output distribution with the pixel decoder's native encoding. On this basis, the pixel stream can then effectively recover the pixel-level details discarded by the semantic encoder's lossy compression within that space. The complete training loss for this stage is:

$$\mathcal{L}_{\text{I}} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{LPIPS}} + \lambda_p \mathcal{L}_p,$$

where $\mathcal{L}_{\text{MSE}}$ represents pixel-wise reconstruction loss, and $\mathcal{L}_{\text{LPIPS}}$ represents the perceptual loss computed using the LPIPS metric.

**Stage II: Joint Decoder Training.** The vision encoder remains frozen while the pixel decoder $\mathcal{D}_p$ is unfrozen for joint training. Only pixel reconstruction losses are used in this stage:

$$\mathcal{L}_{\text{II}} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{LPIPS}}.$$

The decoder adapts to the latent distribution produced by the dual-stream fusion, maximizing reconstruction quality with a fixed encoder.

**Stage III: Encoder Fine-Tuning with Self-Distillation.** The vision encoder is unfrozen for end-to-end training to further enhance pixel representation capability. However, unfreezing the encoder risks catastrophic forgetting: the gradients from reconstruction training may corrupt the discriminative semantic features accumulated during contrastive learning. To address this, we introduce a **semantic alignment loss** that uses the features $f'_v = E'_s(I)$ of a frozen encoder to constrain the fine-tuned encoder features to maintain semantic consistency:

$$\mathcal{L}_s = \|f_v - f'_v\|_2^2.$$

The encoder learning rate is simultaneously decayed to $0.1\times$ the global learning rate, working together with $\mathcal{L}_s$ to form a dual constraint that ensures the encoder

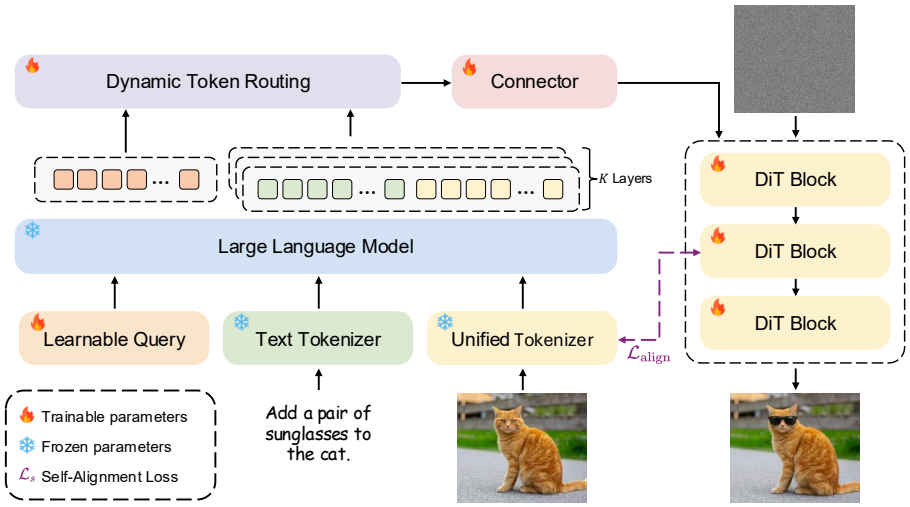



**Fig. 3: Overview of the unified multimodal model.** The frozen MLLM processes multimodal inputs and learnable queries. The DTR adaptively aggregates multi-layer MLLM hidden states based on distinct token semantics to condition the DiT. Furthermore, the optimized tokenizer serves as an internal alignment teacher, establishing a self-alignment paradigm that eliminates reliance on external learners.

does not lose its original semantic understanding performance while acquiring stronger pixel representation capability. Additionally, a GAN discriminator is activated after a certain number of training steps to enhance the perceptual quality of reconstructed images. The complete training loss for this stage is:

$$\mathcal{L}_{\text{III}} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{LPIPS}} + \lambda_s \mathcal{L}_s + \lambda_{\text{GAN}} \mathcal{L}_{\text{GAN}}.$$

### 3.3 Unified Multimodal Model

**Conditional Signal Construction**. To effectively bridge the multimodal reasoning capability of the MLLM with the diffusion-based generation process, we introduce a set of learnable query embeddings as implicit placeholders for the target image representation. During training, specific image generation positions are reserved in the input token sequence, and the query embeddings are inserted at these positions by replacing the corresponding token embeddings as shown in Figure 3. The resulting composite sequence, which contains text tokens, optional reference image tokens, and query embeddings, is then processed uniformly by the MLLM's causal self-attention mechanism. The query positions located toward the end of the sequence naturally aggregate contextual information from the text instructions and visual references, thereby forming conditional representations that fuse multimodal semantics within the MLLM's hidden states.
**Dynamic Token Routing**. Multimodal representations within MLLMs are inherently hierarchical. Existing unified models typically extract only the last-layer

hidden states or employ a token-agnostic static concatenation, completely overlooking the differentiated demands of individual tokens on hierarchical features. To optimally harness these representations for generative guidance, we introduce Dynamic Token Routing (DTR), a mechanism that grants each token the ability to adaptively select its own multi-layer feature fusion weights according to its semantic role. Specifically, the $L$-layer Transformer of the MLLM outputs all $L+1$ hidden states (including the embedding layer output). DTR uniformly samples $K$ layers from them with sampling interval $s = \lfloor L/K \rfloor$ and collects the corresponding hidden states $\{\mathbf{H}^{(k)}\}_{k=1}^{K}$, where $\mathbf{H}^{(k)} \in \mathbb{R}^{B\times N\times D}$, stacked as $\mathbf{H} \in \mathbb{R}^{B\times N\times K\times D}$. The routing network uses the deepest-layer features $\mathbf{H}^{(K)}$ as queries, since the top-layer features contain the most complete semantic context and are best suited for determining the role of each token. It computes the routing weights for the $i$-th token over all layers as:

$$\mathbf{w}_i = \sigma\left(\frac{g(\mathbf{H}_i^{(K)})}{\tau}\right) \in \mathbb{R}^K,$$

where $\sigma(\cdot)$ denotes the softmax function, $g(\cdot)$ is a lightweight routing network, and $\tau$ is a temperature coefficient. The fusion process incorporates learnable per-layer scaling parameters $\alpha \in \mathbb{R}^K$:

$$\hat{\mathbf{H}}_i = \mathbf{W}_p\left(\sum_{k=1}^{K} w_i^{(k)} \cdot \alpha^{(k)} \cdot \mathbf{H}_i^{(k)}\right),$$

where $\mathbf{W}_p$ is a linear projection. The per-token routing weights $\mathbf{w}_i$ produced by DTR also offer a novel interpretability perspective: by visualizing the layer preference distributions of different types of tokens (image edges, texture regions, text instructions), one can intuitively reveal the model's internal decision-making mechanism during generation

**Self-Aligned Generation.** The dual-stream-optimized unified tokenizer possesses both semantic understanding and pixel reconstruction capabilities. We leverage it directly as the representation alignment teacher for the DiT, eliminating the dependence on external pre-trained feature networks [35]. Specifically, we capture the features $f_{\text{dit}}$ at a designated intermediate layer of the DiT. These features are mapped through an alignment projection head $\phi_a$ (MLP) and aligned with the projected features $f'_v$ of the tokenizer's encoder:

$$\mathcal{L}_{\text{align}} = -\frac{1}{N}\sum_{i=1}^{N} \cos\left(\phi_a(f_{\text{dit},i}),\ f'_{v,i}\right),$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity. This self-alignment paradigm seamlessly transfers the semantic and pixel knowledge accumulated during tokenizer training to the generation stage, avoiding the distribution bias that external alignment may introduce while ensuring a consistent semantic feature space shared between the understanding and generation.

**Training Strategy.** The training of the unified model has three stages:

**Table 1:** Comparisons of reconstruction quality on the 256 × 256 ImageNet 50k.

| Model | Ratio | rFID↓ | PSNR↑ | SSIM↑ |
|---|---|---|---|---|
| ***Generative Only Tokenizer*** | | | | |
| LlamaGen [43] | 16 | 2.19 | 20.79 | 0.675 |
| VAR [48] | 16 | 1.00 | 22.63 | 0.755 |
| Open-MAGVIT2 [31] | 16 | 1.67 | 22.70 | 0.640 |
| RAE [82] | 16 | 0.49 | 19.23 | 0.620 |
| SD-VAE [39] | 16 | 2.64 | 22.13 | 0.590 |
| DC-AE [9] | 32 | 0.69 | 23.85 | 0.660 |
| VA-VAE [69] | 16 | 0.28 | **27.96** | 0.790 |
| ***Unified Tokenizer*** | | | | |
| VILA-U [63] | 16 | 1.80 | - | - |
| Tokenflow [37] | 16 | 1.37 | 21.41 | 0.687 |
| DualViTok [17] | 16 | 1.37 | 22.53 | 0.741 |
| DualToken [41] | 16 | 0.54 | 23.56 | 0.742 |
| EMU2 [44] | 14 | 3.27 | 13.49 | 0.420 |
| UniLIP [46] | 32 | 0.79 | 22.99 | 0.747 |
| **SPAR** | 32 | **0.27** | 26.65 | **0.856** |

*Stage I: Connector Pre-training.* The MLLM and DiT are frozen, and only the connector, conditional projection layer, query embeddings, and DTR routing network are trained. Generation data is used with the flow matching loss $\mathcal{L}_{\text{fm}}$ as the training objective. The goal of this stage is to enable the connector to learn the mapping from the MLLM's multimodal output to the DiT's condition space, while allowing the DTR to learn multi-layer fusion weight distribution.

*Stage II: Joint Pre-training.* The MLLM remains frozen while the connector and DiT are jointly trained. The self-alignment loss is added on top of the flow matching loss: $\mathcal{L}_{\text{fm}}+\lambda_a\mathcal{L}_{\text{align}}$. A mixture of generation and editing data is used for training. The DiT releases its parameters in this stage and is co-optimized with the connector, while $\mathcal{L}_{\text{align}}$ transfers the tokenizer's semantic-pixel knowledge to the DiT's intermediate representation space.

*Stage III: Supervised Fine-Tuning.* High-quality instruction tuning datasets are used to further improve generation quality and instruction-following capability. The training configuration remains the same as in Stage II.

## 4 Experiments

### 4.1 Experimental Setups

**Implementation Details.** We implemented two model variants: SPAR-1B and SPAR-3B. SPAR-1B employs InternVL3-1B [85] as the MLLM backbone, which comprises InternViT as the vision encoder and Qwen2.5-0.5B [38] as the language model, paired with SANA-0.6B [65] as the DiT. SPAR-3B adopts InternVL3-2B as the MLLM backbone with a Qwen2.5-1.5B language model and SANA-1.6B DiT. Both variants reuse the InternViT from InternVL3 as the vision encoder

**Table 2:** Comparison with state-of-the-arts on visual understanding benchmarks.

| Model | LLM Params | MME-P | MMB | MMMU | MM-Vet | SEED | MMVP |
|---|---|---|---|---|---|---|---|
| ***Und. Only*** | | | | | | | |
| LLaVA-OV [22] | 1B | 1238 | 52.1 | 31.4 | 29.1 | 65.5 | - |
| InternVL3-1B [85] | 1B | 1492 | 72.6 | 43.4 | 59.5 | 71.1 | 67.3 |
| InternVL3-2B [85] | 2B | 1633 | 80.6 | 48.2 | 62.2 | 75.0 | 72.7 |
| Qwen2.5-VL-3B [2] | 3B | - | 79.1 | 53.1 | 61.8 | - | - |
| Emu3-Chat-8B [54] | 8B | 1244 | 58.5 | 31.6 | 37.2 | 68.2 | 36.6 |
| ***Und. and Gen.*** | | | | | | | |
| Chameleon-7B [47] | 7B | - | 35.7 | 28.4 | 8.3 | - | 0.0 |
| VILA-U-7B [63] | 7B | 1336 | 66.6 | 32.2 | 27.7 | 56.3 | 22.0 |
| MetaMorph-8B [49] | 8B | - | 75.2 | 41.8 | - | - | 48.3 |
| SEED-X-13B [14] | 13B | 1457 | 70.1 | 35.6 | 43.0 | 66.5 | - |
| Show-O-1.3B [66] | 1.3B | 1097 | - | 26.7 | - | - | - |
| Janus-Pro-7B [10] | 7B | 1567 | 79.2 | 41.0 | 50.0 | 72.1 | - |
| Harmon-1.5B [62] | 1.5B | 1155 | 65.5 | 38.9 | - | 67.1 | - |
| BAGEL-7B [11] | 3B | 1610 | 79.2 | 43.2 | 48.2 | - | 54.7 |
| BLIP3-o-4B [7] | 4B | 1528 | 78.6 | 46.6 | 60.1 | 73.8 | - |
| TokLIP-7B [27] | 7B | 1410 | - | 42.1 | - | 65.2 | - |
| Tar-7B [16] | 7B | 1571 | 74.4 | 39.0 | - | 73.0 | - |
| **SPAR-1B** | 1B | 1500 | 73.0 | 43.2 | 59.8 | 71.5 | 68.9 |
| **SPAR-3B** | 2B | **1638** | **80.7** | **48.7** | **62.2** | **75.1** | **73.3** |

and employ the decoder from DC-AE [9] as the pixel decoder. In the asymmetric dual-stream module, the pixel stream uses a 6-layer Transformer encoder, and the semantic stream uses 3 residual blocks. DTR samples 4 layers by default with temperature coefficient $\tau = 1.0$.

**Training Data.** For generation tasks, we used the combination of a 27M re-captioned data publicly released by BLIP3o [7], a 5M subset of CC12M [5], and 4M synthetic images from JourneyDB [42]. For editing tasks, we used the GPT-Image-Edit [58] dataset. In the SFT stage, we employ the BLIP3o-60K and ShareGPT-4o-Image [8] high-quality datasets. Since we froze the LLM throughout training, data for understanding tasks is not required.

### 4.2 Comparisons with SOTA Methods

**Image Reconstruction.** Table 1 presents the comparison of image reconstruction quality on the ImageNet 50k validation set at $256 \times 256$ resolution. Compared to existing unified tokenizers, SPAR achieves state-of-the-art performance. Specifically, our model attains an rFID of 0.27, a PSNR of 26.65, and an SSIM of 0.856, significantly outperforming previous unified methods across all three metrics. Furthermore, compared with generative-only tokenizers, our model still demonstrates highly competitive reconstruction quality, outperforming strong generative baselines such as VA-VAE in SSIM (0.856 vs. 0.790). This indicates that although our tokenizer simultaneously supports both understanding and generation within a unified framework, its reconstruction capability remains competitive with tokenizers designed exclusively for generation. Consequently, it

**Table 3:** Evaluation of text-to-image generation on GenEval and WISE benchmark.

| Model | Params | GenEval | | | WISE | | |
|---|---|---|---|---|---|---|---|
| | | Counting | Position | Overall | Cultural | Biology | Overall |
| ***Gen. Only*** | | | | | | | |
| SDXL [36] | 2.6B | 0.39 | 0.15 | 0.55 | 0.43 | 0.44 | 0.43 |
| FLUX.1-dev [19] | 12B | 0.75 | 0.68 | 0.82 | 0.48 | 0.42 | 0.50 |
| Emu3-Gen [54] | 8B | 0.34 | 0.17 | 0.54 | 0.34 | 0.41 | 0.39 |
| SD3-Medium [12] | 2B | 0.72 | 0.33 | 0.74 | 0.42 | 0.39 | 0.42 |
| Sana-1.6B [65] | 1.6B | 0.62 | 0.21 | 0.66 | - | - | - |
| ***Und. and Gen.*** | | | | | | | |
| VILA-U [63] | 7B | - | - | - | 0.26 | 0.35 | 0.31 |
| TokenFlow-XL [37] | 14B | 0.41 | 0.16 | 0.55 | - | - | - |
| Janus-Pro [10] | 7B | 0.59 | 0.79 | 0.80 | 0.30 | 0.36 | 0.35 |
| Harmon [62] | 3B | 0.66 | 0.74 | 0.76 | 0.38 | 0.37 | 0.41 |
| BLIP3-o-8B [7] | 7B | - | - | 0.84 | - | - | 0.62 |
| BAGEL [11] | 7B | 0.81 | 0.64 | 0.82 | 0.44 | 0.44 | 0.52 |
| OpenUni-B [61] | 1B | 0.74 | 0.77 | 0.84 | 0.37 | 0.39 | 0.43 |
| OpenUni-L [61] | 3B | 0.77 | 0.75 | 0.85 | 0.51 | 0.48 | 0.52 |
| Show-o2 [67] | 7B | 0.58 | 0.52 | 0.76 | 0.33 | 0.39 | 0.39 |
| Tar [16] | 7B | 0.83 | 0.80 | 0.84 | - | - | - |
| **SPAR-1B** | 1B | 0.83 | 0.85 | 0.89 | 0.54 | 0.51 | 0.57 |
| **SPAR-3B** | 3B | **0.84** | **0.87** | **0.91** | **0.67** | **0.61** | **0.64** |

provides a solid foundation for subsequent unified generation models built upon this representation space.

**Multimodal Understanding.** Table 2 presents the comparison of our model with recent advanced methods across various [13, 23, 30, 50, 74, 75] visual understanding benchmarks. In our implementation, we replace the original Vision Encoder of InternVL3 with the ViT from our SPAR tokenizer. Compared to the pure understanding baseline, InternVL3, our model demonstrates superior performance across multiple benchmarks. Notably, although we unfreeze the visual encoder during training, SPAR does not suffer from any performance degradation. On the contrary, benefiting from our dual-stream self-alignment mechanism, the model achieves a noticeable performance enhancement in these understanding tasks. Furthermore, in comparison with contemporary unified models, SPAR consistently achieves superior results across all evaluated metrics.

**Image Generation.** Table 3 reports the text-to-image generation results on the GenEval [15] and WISE [33] benchmarks, and Figure 4 illustrates the qualitative results of our method. We compare SPAR with two categories of methods: generation-only models and unified models that support both understanding and generation. SPAR-3B achieves an overall score of 0.91 on GenEval, ranking first among all compared methods, with the best scores of 0.84 in Counting and 0.87 in Position. Notably, as a unified model, SPAR surpasses even the much larger generation-only model, demonstrating that SPAR's representation space effectively accommodates both understanding and generation without performance degradation. On the knowledge-intensive WISE benchmark, SPAR-3B attains

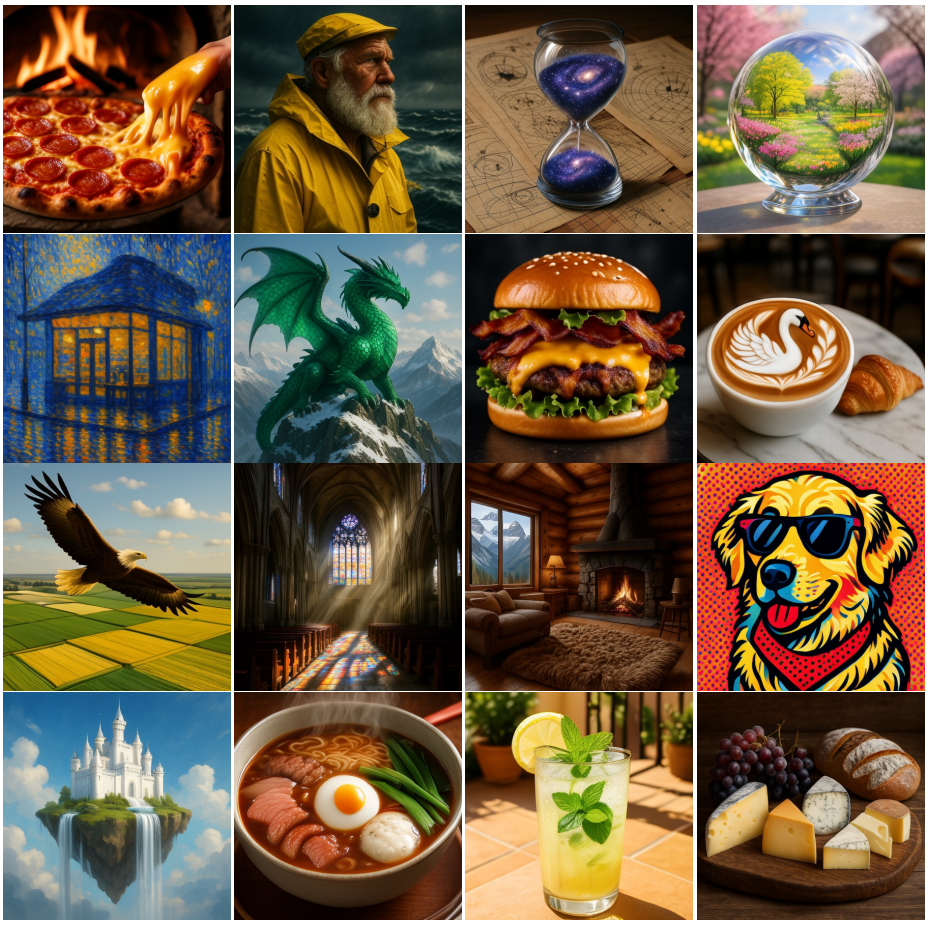


**Fig. 4:** Qualitative results of image generation.

an overall score of 0.64, outperforming all compared methods, further validating that the rich semantic information preserved by the SPAR tokenizer benefits generation quality.

**Image Editing.** We evaluate image editing capabilities on ImgEdit-Bench as shown in Table 4. SPAR-3B achieves an overall score of 4.01, the highest among all open-source methods, closely approaching the proprietary GPT-4o (4.20). Compared with the second-best OmniGen2 (3.44), SPAR-3B yields an absolute improvement of +0.58. Across specific editing categories, SPAR-3B achieves the best performance on various subtypes, with particularly pronounced advantages on semantically demanding tasks, indicating that the rich semantic representations within the SPAR tokenizer effectively enhance the editing model's ability to precisely comprehend and execute editing instructions.

**Table 4:** Evaluation of image editing ability on ImgEdit benchmark.

| Model | Add | Adj. | Ext. | Repl. | Rmv. | Bkg. | Style | Hyb. | Act. | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| GPT-4o [34] | 4.61 | 4.33 | 2.9 | 4.35 | 3.66 | 4.57 | 4.93 | 3.96 | 4.89 | 4.20 |
| MagicBrush [77] | 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 | 1.20 | 1.46 | 1.88 |
| AnyEdit [71] | 3.18 | 2.95 | 1.88 | 2.47 | 2.23 | 2.24 | 2.85 | 1.56 | 2.65 | 2.45 |
| UltraEdit [80] | 3.44 | 2.81 | 2.13 | 2.96 | 1.45 | 2.83 | 3.76 | 1.91 | 2.98 | 2.70 |
| OmniGen [64] | 3.47 | 3.04 | 1.71 | 2.94 | 2.43 | 3.21 | 4.19 | 2.24 | 3.38 | 2.96 |
| Step1X-Edit [29] | 3.88 | 3.14 | 1.76 | 3.40 | 2.41 | 3.16 | 4.63 | 2.64 | 2.52 | 3.06 |
| ICEdit [79] | 3.58 | 3.39 | 1.73 | 3.15 | 2.93 | 3.08 | 3.84 | 2.04 | 3.68 | 3.05 |
| BAGEL [11] | 3.56 | 3.31 | 1.70 | 3.30 | 2.62 | 3.24 | 4.49 | 2.38 | 4.17 | 3.20 |
| UniWorld-V1 [26] | 3.82 | 3.64 | 2.27 | 3.47 | 3.24 | 2.99 | 4.21 | 2.96 | 2.74 | 3.26 |
| OmniGen2 [60] | 3.57 | 3.06 | 1.77 | 3.74 | 3.20 | 3.57 | 4.81 | 2.52 | 4.68 | 3.44 |
| **SPAR-3B** | **4.31** | **3.93** | **2.32** | **4.52** | **4.15** | **4.20** | **4.87** | **3.12** | **4.69** | **4.01** |

**Table 5:** Ablation on unified tokenizer. Evaluated on ImageNet 50k ($256 \times 256$) for reconstruction and multimodal benchmarks for understanding.

| Row | Setting | Reconstruction | | | Understanding | | |
|---|---|---|---|---|---|---|---|
| | | rFID↓ | PSNR↑ | SSIM↑ | MME-P | MMB | MMVP |
| (a) | Full Model | **0.27** | **26.65** | **0.856** | **1500** | **73.0** | **68.9** |
| (b) | w/o Pixel Stream | 0.31 | 24.62 | 0.788 | 1499 | 72.6 | 68.7 |
| (c) | w/o Semantic Stream | 0.29 | 25.28 | 0.804 | 709 | 18.4 | 50.0 |
| (d) | Frozen Encoder | 6.14 | 16.26 | 0.572 | 1492 | 72.6 | 67.3 |

### 4.3 Ablation Studies

**Unified Tokenizer.** Table 5 investigates the contribution of each component in the unified tokenizer. Removing the pixel stream (Row b) leads to a clear reconstruction degradation (PSNR drops by 2.03, SSIM by 0.068), while the understanding metrics remain nearly unchanged, confirming that the semantic encoder alone cannot recover sufficient pixel-level details and the pixel stream is essential for bridging this gap. Removing the semantic stream (Row c) yields better reconstruction than Row b, yet causes a catastrophic collapse in understanding (MMB drops from 73.0 to 18.4, MME-P from 1500 to 709), demonstrating that without the lightweight semantic anchor, end-to-end pixel-oriented optimization destroys the encoder's discriminative structure. The contrast between Rows b and c validates the asymmetric design: the semantic stream preserves understanding at minimal cost, while the heavier pixel stream shoulders the reconstruction burden. Finally, freezing the encoder throughout training (Row d) preserves understanding well but results in severely degraded reconstruction (rFID 6.14, PSNR 16.26), verifying that Stage III encoder fine-tuning with semantic self-distillation is critical for endowing the encoder with pixel representation capability.

**Unified Model.** Table 6 evaluates the two key components of the unified model. Replacing DTR with last-layer-only extraction (Row b) causes the largest overall drop, indicating that adaptively aggregating multi-layer MLLM features provides

**Table 6:** Ablation on unified model training.

| Row | Setting | GenEval↑ | WISE↑ | ImgEdit↑ |
|---|---|---|---|---|
| (a) | Full Model | **0.89** | **0.57** | **3.85** |
| (b) | w/o DTR | 0.86 | 0.54 | 3.73 |
| (c) | w/o Self-Alignment $\mathcal{L}_{\text{align}}$ | 0.86 | 0.53 | 3.78 |

richer structural and semantic cues than relying solely on the top-layer representation. Removing the self-alignment loss $\mathcal{L}_{\text{align}}$ (Row c) also degrades all three metrics, confirming that leveraging the dual-stream tokenizer as an alignment teacher effectively transfers the accumulated semantic-pixel knowledge to the DiT and improves generation quality without any external feature network.

## 5 Conclusion

In this paper, we presented SPAR, a unified multimodal framework that addresses the fundamental feature discrepancy between semantic perception and pixel-level generation. To overcome the detail loss inherent in semantic spaces, we introduced an asymmetric dual-stream tokenizer that explicitly decouples semantic preservation from high-quality pixel reconstruction. Furthermore, we proposed dynamic token routing to adaptively harness multi-layer MLLM representations, along with a self- lignment paradigm that eliminates the reliance on external representation teachers. Extensive experiments demonstrate that SPAR achieves state-of-the-art performance across diverse benchmarks, delivering high-fidelity image generation and complex editing without degrading the pre-trained understanding capabilities.

**Acknowledgment.** This work was supported by National Natural Science Foundation of China (NSFC) Young Scientists Fund Category B (62522216), National Natural Science Foundation of China (NSFC) Young Scientists Fund Category C (62402408), Hong Kong SAR Research Grants Council (RGC) Early Career Scheme (26208924), and Hong Kong SAR Research Grants Council (RGC) General Research Fund (16219025).

## References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 1
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W.,

Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 11
3. Black Forest Labs: FLUX.2: Analyzing and enhancing the latent space of FLUX – representation comparison (2025), https://bfl.ai/research/representation-comparison 3
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023) 14
5. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3558–3568 (2021) 11
6. Chen, H., Li, H., Wang, Z., Chen, L.: Bi-anchor interpolation solver for accelerating generative modeling. arXiv preprint arXiv:2601.21542 (2026) 2
7. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025) 4, 11, 12
8. Chen, J., Cai, Z., Chen, P., Chen, S., Ji, K., Wang, X., Yang, Y., Wang, B.: Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation (2025), https://arxiv.org/abs/2506.18095 4, 11
9. Chen, J., Cai, H., Chen, J., Xie, E., Yang, S., Tang, H., Li, M., Lu, Y., Han, S.: Deep compression autoencoder for efficient high-resolution diffusion models. arXiv preprint arXiv:2410.10733 (2024) 10, 11
10. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025) 11, 12
11. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025) 4, 5, 11, 12, 14
12. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024) 2, 12
13. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023) 12
14. Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., Shan, Y.: Seed-x: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arXiv:2404.14396 (2024) 11
15. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023) 12
16. Han, J., Chen, H., Zhao, Y., Wang, H., Zhao, Q., Yang, Z., He, H., Yue, X., Jiang, L.: Vision as a dialect: Unifying visual understanding and generation via text-aligned representations. arXiv preprint arXiv:2506.18898 (2025) 11, 12
17. Huang, R., Wang, C., Yang, J., Lu, G., Yuan, Y., Han, J., Hou, L., Zhang, W., Hong, L., Zhao, H., et al.: Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. arXiv preprint arXiv:2504.01934 (2025) 10
18. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 2

19. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024) 12
20. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2 (2025) 2
21. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742 2
22. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024) 11
23. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench: Benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13299–13308 (2024) 12
24. Li, H., Li, Y., Lin, B., Niu, Y., Yang, Y., Huang, X., Cai, J., Jiang, X., Hu, Y., Chen, L.: GIR-bench: Versatile benchmark for generating images with reasoning. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=4c1gAsVd9C 4
25. Li, H., Li, Y., Yang, Y., Cao, J., Zhu, Z., Cheng, X., Chen, L.: Dispose: Disentangling pose guidance for controllable human image animation. In: International Conference on Learning Representations. vol. 2025, pp. 72213–72231 (2025) 2
26. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147 (2025) 5, 14
27. Lin, H., Wang, T., Ge, Y., Ge, Y., Lu, Z., Wei, Y., Zhang, Q., Sun, Z., Shan, Y.: Toklip: Marry visual tokens to clip for multimodal comprehension and generation. arXiv preprint arXiv:2505.05422 (2025) 11
28. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023) 1
29. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025) 14
30. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024) 12
31. Luo, Z., Shi, F., Ge, Y., Yang, Y., Wang, L., Shan, Y.: Open-magvit2: An open-source project toward democratizing auto-regressive visual generation (2025), https://arxiv.org/abs/2409.04410 10
32. Ma, C., Jiang, Y., Wu, J., Yang, J., Yu, X., Yuan, Z., Peng, B., Qi, X.: Unitok: A unified tokenizer for visual generation and understanding. arXiv preprint arXiv:2502.20321 (2025) 4
33. Niu, Y., Ning, M., Zheng, M., Jin, W., Lin, B., Jin, P., Liao, J., Ning, K., Feng, C., Zhu, B., Yuan, L.: Wise: A world knowledge-informed semantic evaluation for text-to-image generation. arXiv preprint arXiv:2503.07265 (2025) 12
34. OpenAI: Introducing 4o Image Generation (2025), https://openai.com/index/introducing-4o-image-generation/ 14
35. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve,

G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023) 2, 3, 4, 9
36. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 12
37. Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: Tokenflow: Unified image tokenizer for multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2545–2555 (2025) 4, 10, 12
38. Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report (2025), https://arxiv.org/abs/2412.15115 10
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2, 10
40. Shi, M., Wang, H., Zheng, W., Yuan, Z., Wu, X., Wang, X., Wan, P., Zhou, J., Lu, J.: Latent diffusion model without variational autoencoder (2025), https://arxiv.org/abs/2510.15301 4
41. Song, W., Wang, Y., Song, Z., Li, Y., Sun, H., Chen, W., Zhou, Z., Xu, J., Wang, J., Yu, K.: Dualtoken: Towards unifying visual understanding and generation with dual visual vocabularies. arXiv preprint arXiv:2503.14324 (2025) 10
42. Sun, K., Pan, J., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., et al.: Journeydb: A benchmark for generative image understanding. Advances in neural information processing systems **36**, 49659–49678 (2023) 11
43. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024) 10
44. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14398–14409 (2024) 4, 10
45. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Emu: Generative pretraining in multimodality. arXiv preprint arXiv:2307.05222 (2023) 4
46. Tang, H., Xie, C., Bao, X., Weng, T., Li, P., Zheng, Y., Wang, L.: Unilip: Adapting clip for unified multimodal understanding, generation and editing. arXiv preprint arXiv:2507.23278 (2025) 4, 5, 10
47. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024) 4, 11
48. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024) 2, 10
49. Tong, S., Fan, D., Zhu, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. arXiv preprint arXiv:2412.14164 (2024) 11

50. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9568–9578 (2024) 12
51. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 2
52. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017) 4
53. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025) 1
54. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024) 11, 12
55. Wang, Y., Chen, H., Liu, J., He, Z., Liu, R., Wang, Z., Chen, L.: Lisa: Likelihood score alignment for visual-condition controllable generation. arXiv preprint arXiv:2606.27192 (2026) 2
56. Wang, Y., Jiang, Z., Wang, Z., Chen, L.: Coarse-guided visual generation via weighted h-transform sampling. arXiv preprint arXiv:2603.12057 (2026) 2
57. Wang, Y., Wang, Z., Chen, L.: Target-aware image editing via cycle-consistent constraints. arXiv preprint arXiv:2510.20212 (2025) 2
58. Wang, Y., Yang, S., Zhao, B., Zhang, L., Liu, Q., Zhou, Y., Xie, C.: Gpt-image-edit-1.5 m: A million-scale, gpt-generated image dataset. arXiv preprint arXiv:2507.21033 (2025) 11
59. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025) 2
60. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025) 14
61. Wu, S., Wu, Z., Gong, Z., Tao, Q., Jin, S., Li, Q., Li, W., Loy, C.C.: Openuni: A simple baseline for unified multimodal understanding and generation. arXiv preprint arXiv:2505.23661 (2025) 4, 12
62. Wu, S., Zhang, W., Xu, L., Jin, S., Wu, Z., Tao, Q., Liu, W., Li, W., Loy, C.C.: Harmonizing visual representations for unified multimodal understanding and generation (2025), https://arxiv.org/abs/2503.21979 11, 12
63. Wu, Y., Zhang, Z., Chen, J., Tang, H., Li, D., Fang, Y., Zhu, L., Xie, E., Yin, H., Yi, L., et al.: Vila-u: a unified foundation model integrating visual understanding and generation. arXiv preprint arXiv:2409.04429 (2024) 4, 10, 11, 12
64. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13294–13304 (2025) 14
65. Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., Han, S.: Sana: Efficient high-resolution image synthesis with linear diffusion transformer (2024), https://arxiv.org/abs/2410.10629 10, 12
66. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024) 4, 11
67. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. arXiv preprint arXiv:2506.15564 (2025) 12

68. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 2
69. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15703–15712 (2025) 2, 3, 10
70. Yu, L., Shi, B., Pasunuru, R., Muller, B., Golovneva, O., Wang, T., Babu, A., Tang, B., Karrer, B., Sheynin, S., Ross, C., Polyak, A., Howes, R., Sharma, V., Xu, P., Tamoyan, H., Ashual, O., Singer, U., Li, S.W., Zhang, S., James, R., Ghosh, G., Taigman, Y., Fazel-Zarandi, M., Celikyilmaz, A., Zettlemoyer, L., Aghajanyan, A.: Scaling autoregressive multi-modal models: Pretraining and instruction tuning (2023), https://arxiv.org/abs/2309.02591 4
71. Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., Zhuang, Y.: Anyedit: Mastering unified high-quality image editing for any idea. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26125–26135 (2025) 14
72. Yu, Q., Weber, M., Deng, X., Shen, X., Cremers, D., Chen, L.C.: An image is worth 32 tokens for reconstruction and generation. Advances in Neural Information Processing Systems **37**, 128940–128966 (2024) 2
73. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. In: International Conference on Learning Representations (2025) 2, 3, 4
74. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023) 12
75. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9556–9567 (2024) 12
76. Yue, Z., Zhang, H., Zeng, X., Chen, B., Wang, C., Zhuang, S., Dong, L., Du, K., Wang, Y., Wang, L., et al.: Uniflow: A unified pixel flow tokenizer for visual understanding and generation. arXiv preprint arXiv:2510.10575 (2025) 4
77. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36**, 31428–31449 (2023) 14
78. Zhang, S., Zhang, H., Zhang, Z., Ge, C., Xue, S., Liu, S., Ren, M., Kim, S.Y., Zhou, Y., Liu, Q., et al.: Both semantics and reconstruction matter: Making representation encoders ready for text-to-image generation and editing. arXiv preprint arXiv:2512.17909 (2025) 3, 4
79. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: Enabling instructional image editing with in-context generation in large scale diffusion transformer. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025) 14
80. Zhao, H., Ma, X.S., Chen, L., Si, S., Wu, R., An, K., Yu, P., Zhang, M., Li, Q., Chang, B.: Ultraedit: Instruction-based fine-grained image editing at scale. Advances in Neural Information Processing Systems **37**, 3058–3093 (2024) 14

81. Zhao, S., Zhang, X., Guo, J., Hu, J., Duan, L., Fu, M., Chng, Y.X., Wang, G.H., Chen, Q.G., Xu, Z., et al.: Unified multimodal understanding and generation models: Advances, challenges, and opportunities. arXiv preprint arXiv:2505.02567 (2025) 4
82. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025) 2, 4, 10
83. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039 (2024) 4
84. Zhu, C., Li, H., Li, X., Chen, L.: Mokus: Leveraging cross-modal knowledge transfer for knowledge-aware concept customization. arXiv preprint arXiv:2603.12743 (2026) 4
85. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025) 10, 11