# CHEERS: Decoupling Patch Details from Semantic Representations Enables Unified Multimodal Comprehension and Generation

Yichen Zhang[1*], Da Peng[2*], Zonghao Guo[1†], Zijian Zhang[3], Xuesong Yang[3], Tong Sun[3], Shichu Sun[3], Yidan Zhang[3], Yanghao Li[1], Haiyan Zhao[1], Wang Xu[1], Qi Shi[1], Yangang Sun[1], Chi Chen[1], Shuo Wang[1], Yukun Yan[1], Xu Han[1], Qiang Ma[1], Wei Ke[2], Liang Wang[1], Zhiyuan Liu[1], and Maosong Sun[1]
{yichen0zhang, metapda}@gmail.com, guozonghao96@outlook.com

[1] Tsinghua University, Beijing, China
[2] Xi'an Jiaotong University, Xi'an, China
[3] University of Chinese Academy of Sciences, Beijing, China

**Abstract.** A recent cutting-edge topic in multimodal modeling is to unify visual comprehension and generation within a single model. However, the two tasks demand mismatched decoding regimes and visual representations, making it non-trivial to jointly optimize within a shared feature space. In this work, we present CHEERS, a unified multimodal model that decouples patch-level details from semantic representations, thereby stabilizing semantics for multimodal understanding and improving fidelity for image generation via gated detail residuals. CHEERS includes three key components: (i) a unified vision tokenizer that encodes and compresses image latent states into semantic tokens for efficient LLM conditioning, (ii) an LLM-based Transformer that unifies autoregressive decoding for text generation and diffusion decoding for image generation, and (iii) a cascaded flow matching head that decodes visual semantics first and then injects semantically gated detail residuals from the vision tokenizer to refine high-frequency content. Experiments on popular benchmarks demonstrate that CHEERS matches or surpasses advanced UMMs in both visual understanding and generation. Notably, CHEERS outperforms the Tar-1.5B on the popular benchmarks GenEval and MMBench, while requiring only 20% of the training cost, indicating effective and efficient (*i.e.*, 4× token compression) unified multimodal modeling. Code and checkpoints are available at github.com/AI9Stars/CHEERS.

**Keywords:** Unified multimodal model · Visual generation and comprehension · Unified vision encoder.

## 1 Introduction

Multimodal large language models (MLLMs) [3, 43, 60, 71, 88] have largely matured for visual comprehension, while diffusion models [25, 26, 32, 42, 47, 90] have

* Equal contribution. † Corresponding author.
This work is supported by the AI9Stars community.

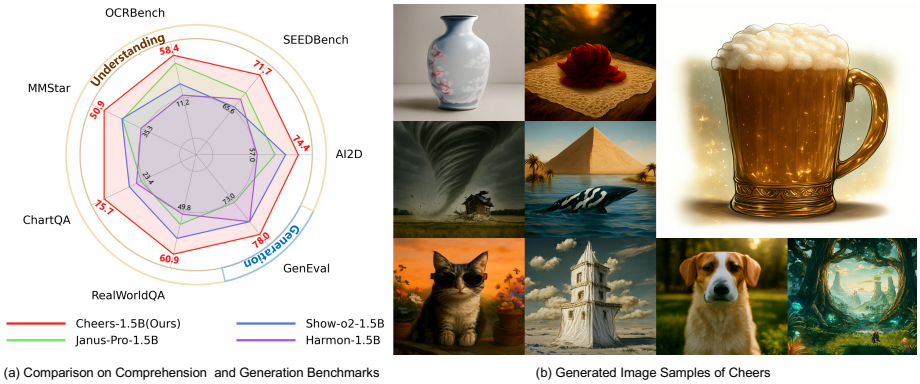


**Fig. 1:** Cheers Capabilities. (a) Performance on general understanding and generation benchmarks compared with unified multimodal models (UMMs) of similar scale. (b) Generated image samples of Cheers.

set the standard for high-fidelity image generation. Bringing both into a single model is a cutting-edge step toward more human-like multimodal intelligence. However, such unification is particularly challenging, as the two tasks demand fundamentally different decoding mechanisms and visual representations.

In terms of decoding mechanisms, discretizing visual representations [19, 21, 63, 66] for autoregressive (AR) prediction with text tokens offers a seamless adaptation to existing MLLM architectures [33, 52, 91]. However, discrete tokens suffer from quantization errors [59, 78] and dimensional constraints [52, 59, 89], leading to the loss of visual information. Bypassing the constraints of sequential raster-scanning of image generation, recent approaches [12, 36, 80] integrate diffusion modeling to capture global visual context alongside AR-based text generation.

From the perspective of visual representations, multimodal understanding typically relies on semantic-rich features from vision encoders [46, 65], whereas high-fidelity image generation often depends on detail-preserving latents from reconstruction-oriented tokenizers [24, 66, 83]. However, relying solely on a single representation often fails to simultaneously satisfy these distinct requirements [11, 55, 56, 59, 61, 72, 79], as shown in Fig. 2 (b). Therefore, one line of UMMs [9, 12, 74] separates the feature optimization for visual comprehension and generation, achieving strong task-specific performance, as shown in Fig. 2 (a). The other line seeks to integrate these capabilities via a unified token interface by either fusing heterogeneous features [30, 80] or jointly optimizing a shared vision tokenizer with multiple objectives [13, 45, 76, 82], as shown in Fig. 2 (c).

Despite these inspiring explorations, the intrinsic optimization conflict between visual comprehension and generation remains insufficiently investigated in UMMs. In this paper, we introduce Cheers, a UMM that decouples patch-level details from semantic representations, stabilizing semantics for image understanding and improving generation fidelity by injecting high-frequency detail residuals, as shown in Fig. 2 (d). Cheers includes three key components. (i)

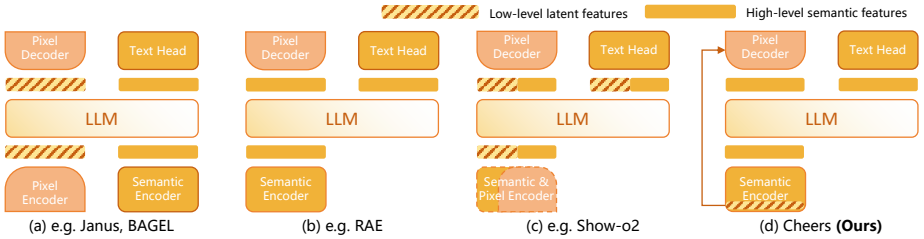


**Fig. 2:** Architectural comparison between prior UMMs and CHEERS. (a) Separated visual spaces for understanding and generation. (b) Single semantic-centric space with limited structural details. (c) Fused feature representation with potential interference. (d) Cheers (Ours): A unified vision tokenizer that integrates structural and semantic features to ensure stable semantic understanding while enhancing generative details.

A unified vision tokenizer utilizes a representation encoder upon VAE latents to extract semantic features, subsequently compressed via a pixel-unshuffle [71] operation for efficient LLM conditioning. (ii) An LLM-based Transformer integrates autoregressive and diffusion decoding for text and image generation, respectively, thereby capitalizing on the superior modeling paradigms inherent to each modality. (iii) Cascaded Flow Matching Head decouples image generation into two phases. It first synthesizes low-resolution global semantics, then injects semantically gated high-frequency residuals to achieve precise super-resolution, adopting a coarse-to-fine approach that resonates with recent work [2].

Extensive experiments on standard benchmarks demonstrate that CHEERS performs on par with or exceeds state-of-the-art UMMs in both visual comprehension and generation, validating the efficacy of our unified modeling approach. CHEERS also represents a significant step towards token-compressed UMMs, achieving a 4× compression rate for efficient high-resolution image understanding and generation. Notably, CHEERS outperforms the Tar model on the popular benchmarks GenEval and MMBench, while requiring only 20% of the training cost, indicating effective and efficient unified multimodal modeling.

Our contribution can be summarized as threefold. (1) We propose decoupling patch details from semantic representations, which redefine the multimodal feature modeling trajectory of UMMs, alleviating the optimization interference between comprehension and generation tasks. (2) We introduce CHEERS, a hybrid-decoding UMM equipped with a unified vision tokenizer that achieves significant token compression for efficient multimodal modeling. (3) We perform extensive evaluations on popular benchmarks to verify the effectiveness of CHEERS, providing detailed analysis and insights for future research.

## 2 CHEERS

We present the CHEERS framework, covering its architecture (Sec. 2.1), inference strategy and objectives (Sec. 2.2), and training pipeline (Sec. 2.3).

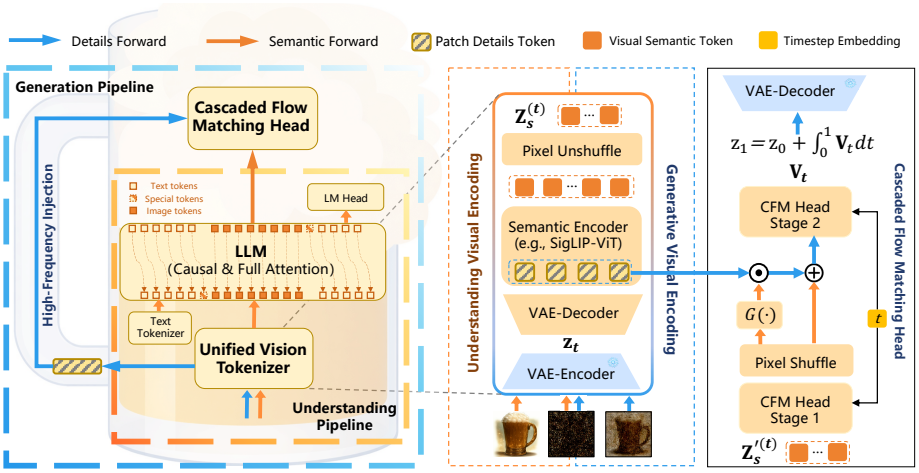


**Fig. 3:** Overview of CHEERS, a unified framework for multimodal understanding and image generation. The Unified Vision Tokenizer converts visual inputs into semantic tokens that are jointly processed with text tokens by the LLM for understanding tasks, and detail tokens that serve as step-adaptive high-frequency injection into the CFM Head during generation. During generation, the CFM Head predicts a continuous-time velocity field in the latent space, enabling iterative sampling from Gaussian noise $\mathbf{z}_0$ to the terminal latent $\mathbf{z}_1$, which is finally decoded by the VAE decoder.

## 2.1 Model Architecture

As illustrated in Fig. 3, CHEERS is built upon three key components: a unified vision tokenizer for visual encoding, a unified LLM-based Transformer backbone for multimodal modeling, and a cascaded flow matching head for image generation. Additionally, a standard text tokenizer and a language modeling (LM) head are employed for language encoding and text generation, respectively, to support visual understanding tasks.

**Unified Vision Tokenizer.** As illustrated in Fig. 3, CHEERS adopts a unified vision tokenizer composed of a VAE decoder [25] and a semantic encoder, *i.e.*, SigLIP2-ViT [65]. Specifically, the latent representations produced by the VAE encoder are first decoded into the image space via the VAE decoder, after which SigLIP2 extracts high-level semantic visual features. In this way, the VAE decoder and SigLIP2 jointly function as an integrated module that bridges latent representations and unified semantic visual embeddings.

Specifically, given an input image $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$, where $H$ and $W$ denote the image height and width, we first process it through a VAE encoder, yielding the latent states $\mathbf{z}_1 \in \mathbb{R}^{h\times w\times d}$, where $h = H/16$, $w = W/16$, and $d$ represents the latent feature dimension. To unify diverse tasks, we formulate a task-dependent latent $\mathbf{z}_t = t\mathbf{z}_1 + (1-t)\mathbf{z}_0$, where latent noise $\mathbf{z}_0 \sim \mathcal{N}(0,1)$. We sample timestep $t \in (0,1)$ for image generation, fix $t = 1$ (*i.e.*, $\mathbf{z}_t = \mathbf{z}_1$) for visual understanding, and set $t = 0$ (*i.e.*, $\mathbf{z}_t = \mathbf{z}_0$) for language-only tasks. Subsequently, instead of

directly processing these latent states $\mathbf{z}_t$ using a ViT with randomly initialized patch embeddings like [36, 80], $\mathbf{z}_t$ is passed through a VAE decoder $D(\cdot)$ to reconstruct the pixel-level image. The reconstructed image is then encoded by the ViT backbone to extract high-level semantic tokens $\mathbf{z}_s^{(t)} \in \mathbb{R}^{h\times w\times d'}$, where $d'$ denotes the semantic feature dimension. To ensure strict spatial alignment between these semantic tokens and the latent patches, we adopt SigLIP2-ViT with a 16×16 patch embedding layer $S(\cdot)$. Notably, we experimentally found that direct latent processing like [36] discards fine-grained features and hinders OCR-centric understanding ability. By reconstructing the pixel space, we circumvent this issue and successfully retain essential visual details. Please kindly refer to the Supplement for details.

Before feeding the semantic tokens into our unified LLM-based Transformer, a Pixel-Unshuffle module [71] is applied to reduce their spatial resolution and project the channel dimension, resulting in $\mathbf{Z}_s^{(t)} \in \mathbb{R}^{h/2\times w/2\times c}$, where $c$ is the LLM hidden size. To the best of our knowledge, we are the first work to introduce 2D token compression within a UMM.

**Unified LLM-based Transformer.** To achieve optimal image-text joint modeling, we utilize autoregressive decoding for text generation and diffusion processes for image generation within a single LLM backbone, *i.e.*, Qwen2.5-1.5B-Instruct [62]. Specifically, given the semantic visual tokens $\mathbf{Z}_s^{(t)}$ and the text embeddings $\mathbf{Z}_{text}$ derived from input instructions via the text tokenizer, we concatenate them into a unified input sequence, which is then processed by the LLM backbone to yield contextualized hidden states through deep cross-modal encoding. Note that a bidirectional attention mask is applied to $\mathbf{Z}_s^{(t)}$ to capture global visual context, whereas a causal mask is employed for $\mathbf{Z}_{text}$ to enable AR decoding. Depending on the task modality, the LLM outputs are subsequently routed to different decoding paradigms. For visual comprehension or pure text generation, the model employs a standard AR language modeling objective. For image generation, the continuous visual hidden states $\mathbf{Z}_s^{(t)}$, which have been integrated with the text instructions or descriptions $\mathbf{Z}_{text}$, are decoded via our cascaded flow matching head.

**Cascaded Flow Matching Head.** Inspired by [2, 5, 15], we propose to explicitly decouple high-frequency visual details from low-frequency semantic features and then integrate them during image synthesis. Specifically, our CFM head consists of two cascaded stages, comprising 7 and 3 DiT blocks [42] respectively. Both stages employ the AdaLN-Zero [42] architecture to incorporate temporal modulations of the denoising procedure from the timestep $t$. In the first stage, the CFM head takes the contextualized hidden states $\mathbf{Z}_s^{(t)} \in \mathbb{R}^{h/2\times w/2\times c}$ from the LLM as input to perform low-resolution semantic generation. This is followed by a PixelShuffle [50] module that up-samples the feature maps to 2× resolution and low-dimension ones $\mathbf{Z'}_s^{(t)} \in \mathbb{R}^{h\times w\times d'}$. In the second stage, given the high-frequency patch details $S(D(\mathbf{z}_t)) \in \mathbb{R}^{h\times w\times d'}$, we first introduce a gating network $G(\cdot)$ to adaptively control the injection of fine-grained information to

update the decoded features $\mathbf{Z'}_s^{(t)}$ as

$$\mathbf{Z'}_s^{(t)} \leftarrow G(\mathbf{Z'}_s^{(t)}) \odot S(D(\mathbf{z}_t)) + \mathbf{Z'}_s^{(t)},$$

where $G(\mathbf{Z'}_s^{(t)}) \in \mathbb{R}^{h \times w \times 1}$ denotes a scalar map and $\odot$ the element-wise multiplication. Notably, as $\mathbf{Z'}_s^{(t)}$ is modulated by the timestep $t$ in the first stage, the intensity of high-frequency injection (HFI) is dynamically coupled with the generative trajectory. Our empirical analysis (see Sec. 3.3) reveals that, even without explicit supervision, the magnitude of HFI naturally intensifies as $t$ progresses.

Finally, $\mathbf{Z'}_s^{(t)}$ is fed into subsequent DiT layers to predict the velocity field $\mathbf{V}_t$. Such a progression mirrors the hierarchical nature of human drawing, which naturally transitions from global layout sketching to localized detail refinement.

### 2.2 Inference and Training Objectives

**Inference.** For text-only and multimodal understanding tasks, we follow standard autoregressive decoding by sequentially selecting tokens from the predicted distribution. For image generation, we perform continuous-time flow-based sampling starting from Gaussian noise in the latent space, denoted as $\mathbf{z}_0$. At each time step $t$, we feed the current latent variable $\mathbf{z}_t$ into the unified vision tokenizer to obtain the corresponding visual tokens, which are then jointly processed with the textual condition by the LLM. Subsequently, the CFM head predicts the continuous-time velocity field $\mathbf{V}_t$ based on the LLM outputs, and we update the latent state via numerical integration:

$$\mathbf{z}_{t+\Delta t} = \mathbf{z}_t + \int_t^{t+\Delta t} \mathbf{V}_\tau \, d\tau.$$

The updated latent variable $\mathbf{z}_{t+\Delta t}$ serves as the input to the next integration step. By repeatedly applying tokenization, conditional modeling with the LLM, velocity prediction through the CFM Head, and numerical ODE integration, the latent trajectory is evolved from $\mathbf{z}_0$ to the terminal state $\mathbf{z}_1$. The final latent $\mathbf{z}_1$ is then decoded using the VAE decoder to produce the output image.

In addition, following prior work [80], we adopt classifier-free guidance (CFG) during generation. To further adjust the time noise schedule in flow-based sampling, we apply a schedule shift and rescale the continuous-time variable with a hyperparameter $\alpha$. Formally, given the original time step $t \in [0, 1]$, the shifted time step is computed as $\tilde{t} = \frac{\alpha t}{1+(\alpha-1)t}$.

**Training Objectives.** We use an end-to-end unified training optimization. For visual comprehension or pure text generation, the probability of generating the target text sequence $\mathbf{y} = \{y_1, \ldots, y_L\}$ is factorized as $P_\theta(\mathbf{y}|\mathbf{C}) = \prod_{i=1}^{L} p_\theta(y_i|\mathbf{y}_{<i}, \mathbf{C})$, where $\mathbf{C}$ represents the conditioning context, *i.e.*, $[\mathbf{Z}_s^{(t)}]$ for image caption, $[\mathbf{Z}_s^{(t)}, \mathbf{Z}_{text}]$ for image question-answer or prefix $[\mathbf{Z}_{text}]$ for pure text. We use the standard cross-entropy loss function $\mathcal{L}_{AR} = -\log P_\theta(\mathbf{y}|\mathbf{C})$, where $\mathbf{y}$ is the generated target text sequence, $\mathbf{C}$ is the conditioning context,

and $\theta$ are the learnable parameters of the model. For the image generation part, we use the flow matching loss function $\mathcal{L}_{FM} = \|v_\theta(\mathbf{Z}'^{(t)}_s) - (\mathbf{z}_1 - \mathbf{z}_0)\|_2^2$. The overall training loss is the weighted sum of the text loss and image generation loss, given by:

$$\mathcal{L}_{total} = \mathcal{L}_{AR} + \lambda \mathcal{L}_{FM}$$

where $\lambda$ is a hyperparameter used to balance the loss between text generation and image generation, and in our training, $\lambda$ is set to 1.

### 2.3 Training Pipeline

Our four-stage progressive training is detailed in Tab. 1. Image resolution is fixed at $512 \times 512$. We initialize the image encoder from Siglip2 and FLUX.2. All experiments use the AdamW optimizer with a 0.02 warmup ratio and 1.0 gradient clipping, conducted on 128 NVIDIA A100 GPUs (16 nodes).

**Stage I: Vision–Language Alignment.** We train only the randomly initialized modules (projector, CFM head, and gating modules). The training data consists of 4.5M image-caption pairs from the LLaVA-UHD-v3 [57] and 1.3M ImageNet samples re-annotated by Qwen2.5-VL-3B [3]. To establish preliminary generative capability, we repeat the ImageNet dataset 10 times.

**Stage II: General Pre-Training.** Subsequently, we optimize all model parameters except the VAE using 30M multimodal samples. Understanding data comprises captions from Infinity-MM [18], LLaVA-UHD-v3 [57], and TextAtlas5M [67]. Generation data, including pretraining data from BLIP-3o [6], and a small portion of synthetic data re-generated using FLUX.2-klein-9B [25] with prompts from DiffusionDB [73]. Pure text data extracted from LLaVA-UHD-v3 [57]. The ratio of understanding, generation, and text data is $3 : 6 : 1$.

**Stage III: Refined Pre-Training.** We focus on visual reasoning and semantic alignment using 33M samples in this stage, maintaining a $3 : 6 : 1$ ratio across understanding, generation, and text data. We combine LLaVA-UHD-v3 instruction data [57] for understanding and synthetic data generated via FLUX.2-klein-9B [25], utilizing prompts from DiffusionDB [73] and LLaVA-OneVision-1.5 [1]. To improve compositional reasoning (*e.g.*, counting, color, and space), we also produced 466K instructions based on Objects365 [48] to synthesize images. Pure text data is extracted from Nemotron-Cascade [68].

**Table 1:** Training setup and hyperparameters across different stages for CHEERS.

| Stage | Vision-Language Alignment | General Pre-Training | Refined Pre-Training | Supervised Fine-Tuning |
|---|---|---|---|---|
| Dataset | Image caption<br>Text-to-image | Pure text corpus<br>Detailed image caption<br>Text-to-image<br>OCR data | Pure text corpus<br>Synthetic generation data<br>General VQA | Pure text corpus<br>Instruction text-to-image<br>High-quality instruction data |
| Learning rate | $1e^{-4}$ | $1e^{-4}$ | $4e^{-5}$ | $2e^{-5}$ |
| Training parts | Projection & CFM Head & Gate | All parameters w/o VAE | All parameters w/o VAE | All parameters w/o VAE |
| Schedules | constant | constant | constant | cosine |
| Batch size | 512 | 512 | 512 | 128 |
| Training steps | 30 K | 60 K | 65 K | 30 K |

**Stage IV: Supervised Fine-Tuning.** We fine-tune the model on 3.8M curated samples, incorporating a high-quality subset of the Stage III data with Echo-4o-Image [84], MoviePosters [51], and ShareGPT-4o-Image [7]. During training, we maintain a 1 : 1 batch ratio between understanding and generation tasks.

## 3 Experiments

We evaluate CHEERS on diverse multimodal benchmarks. We first describe the setup in Sec. 3.1 and report main results in Sec. 3.2. Subsequent analyses include visualizations in Sec. 3.3, ablation studies in Sec. 3.4, and discussion of the model's characteristics in Sec. 3.5.

### 3.1 Evaluation Setup

**Multimodal Understanding.** We evaluate CHEERS on diverse and widely recognized multimodal understanding benchmarks. (1) General Benchmarks: SEEDBench [27], MMStar [8], MMBench [34]. (2) OCR Benchmarks: ChartQA [41], OCRBench [35]. (3) Visual Spatial Benchmarks: RealWorldQA [77], POPE [28]. (4) Knowledge-focused Benchmarks: AI2D [23], MathVista [38], MMMU [87].
**Visual Generation.** We evaluate visual generation on GenEval [17] and DPG-Bench [22]. GenEval assesses compositional alignment and fine-grained controllable generation. DPG-Bench tests semantic alignment and prompt-following in complex, multi-entity scenarios using dense prompts.

### 3.2 Main Results

**Image Understanding.** As shown in Tab. 2, CHEERS achieves competitive performance on nearly all benchmarks, demonstrating strong understanding ability.
**Image Generation.** The results are summarized in Tab. 3 and Tab. 4. Across all benchmarks, CHEERS consistently achieves competitive or superior performance compared with existing approaches under comparable parameter scales, including models such as Janus-Pro. Notably, CHEERS attains these strong results using only 83M training samples in total, demonstrating that high-quality

**Table 2:** Multimodal understanding benchmarks. #Params.: LLM backbone size.

| Model | #Params. | General | | | OCR | | Visual Spatial | | Knowledge | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | SEEDBench | MMStar | MMBench | ChartQA | OCRBench | RealWorldQA | POPE | AI2D | MathVista | MMMU |
| ***Understanding Only*** | | | | | | | | | | | |
| MobileVLM-V2 [10] | 1.4B | - | - | 57.7 | - | - | - | 84.3 | - | - | - |
| Qwen2-VL [70] | 2B | - | 48.0 | 72.2 | 73.5 | 80.9 | 62.9 | - | 74.7 | 43.0 | 41.1 |
| DeepSeek-VL [37] | 7B | 70.4 | - | 73.2 | - | 45.6 | - | 88.1 | - | 36.1 | 36.6 |
| mPLUG-Owl2 [85] | 7B | 57.8 | - | 64.5 | 22.8 | 25.5 | 50.3 | 86.2 | 55.7 | - | 32.7 |
| ***Understanding & Generation*** | | | | | | | | | | | |
| Emu3 [72] | 8B | 68.2 | - | 58.5 | <u>68.6</u> | **68.7** | <u>57.4</u> | 85.2 | <u>70.0</u> | - | 31.6 |
| Show-o [79] | 1.3B | 51.5 | - | - | - | - | - | 80.0 | - | - | 26.7 |
| Show-o2 [80] | 1.5B | 65.6 | <u>43.4</u> | 67.4 | 40.0 | 24.5 | 56.5 | - | 69.0 | - | <u>37.1</u> |
| JanusFlow [40] | 1.3B | <u>70.5</u> | 40.6 | <u>74.9</u> | 64.6 | 53.2 | 41.2 | <u>88.0</u> | 54.2 | - | 29.3 |
| Janus-Pro [9] | 1.5B | 68.3 | 43.1 | **75.5** | 23.4 | 48.7 | 52.6 | 86.2 | 64.5 | - | 36.3 |
| Harmon [75] | 1.5B | 67.1 | 35.3 | 65.5 | 29.8 | 11.2 | 49.8 | 87.6 | 57.0 | - | **38.9** |
| Tar [20] | 1.5B | 70.4 | - | 65.6 | - | - | - | **88.4** | - | - | 36.0 |
| CHEERS | 1.5B | **71.7** | **50.9** | 70.4 | **75.7** | <u>58.4</u> | **60.9** | 87.9 | **74.4** | **50.5** | 36.0 |

**Table 3:** Performances on GenEval. Obj.: Object; Attr.: Attribute; #Params.: LLM backbone parameters; #Data: total training samples (visual, image, and text).

| Model | #Params. | #Data | Single Obj. | Two Obj. | Counting | Colors | Position | Color Attr. | Overall |
|---|---|---|---|---|---|---|---|---|---|
| ***Generation Only*** | | | | | | | | | |
| LlamaGen [54] | 0.8B | 62M | 0.71 | 0.34 | 0.21 | 0.58 | 0.07 | 0.04 | 0.32 |
| SDXL [44] | 2.6B | - | 0.98 | 0.74 | 0.39 | 0.85 | 0.15 | 0.23 | 0.55 |
| DALL-E 3 [4] | - | - | 0.96 | 0.87 | 0.47 | 0.83 | 0.43 | 0.45 | 0.67 |
| SD3-Medium [14] | 2B | - | 0.99 | 0.94 | 0.72 | 0.89 | 0.33 | 0.60 | 0.74 |
| ***Understanding & Generation*** | | | | | | | | | |
| Chameleon [59] | 7B | 5.2B | - | - | - | - | - | - | 0.39 |
| Show-o [79] | 1.3B | 3.2B | 0.95 | 0.52 | 0.49 | 0.82 | 0.11 | 0.28 | 0.53 |
| TokenFlow-XL [45] | 14B | 5.76B | 0.95 | 0.60 | 0.41 | 0.81 | 0.16 | 0.24 | 0.55 |
| Janus [74] | 1.3B | - | 0.97 | 0.68 | 0.30 | 0.84 | 0.46 | 0.42 | 0.61 |
| JanusFlow [40] | 1.3B | - | - | - | - | - | - | - | 0.63 |
| Transfusion [91] | 7.3B | 3.5B | - | - | - | - | - | - | 0.63 |
| D-DiT [29] | 2B | 40M | 0.97 | 0.80 | 0.54 | 0.76 | 0.32 | 0.50 | 0.65 |
| Emu3 [72] | 8B | - | 0.99 | 0.81 | 0.42 | 0.80 | 0.49 | 0.45 | 0.66 |
| Janus-Pro [9] | 1.5B | 162M | 0.98 | 0.82 | 0.51 | 0.89 | 0.65 | 0.56 | 0.73 |
| Show-o2 [80] | 1.5B | 177M | 0.99 | 0.86 | 0.55 | 0.86 | 0.46 | 0.63 | 0.73 |
| Harmon [75] | 1.5B | 113M | 0.99 | 0.86 | 0.66 | 0.85 | 0.74 | 0.48 | <u>0.76</u> |
| Tar [20] | 1.5B | 403M | 0.99 | 0.91 | 0.76 | 0.81 | 0.57 | 0.51 | <u>0.76</u> |
| CHEERS | 1.5B | 83M | 0.98 | 0.92 | 0.65 | 0.86 | 0.63 | 0.65 | **0.78** |

**Table 4:** Performances on DPG-Bench. #Params.: LLM backbone parameters; #Data: total training samples (visual, image, and text).

| Model | #Params. | #Data | Global | Entity | Attribute | Relation | Other | Overall |
|---|---|---|---|---|---|---|---|---|
| ***Generation Only*** | | | | | | | | |
| SDXL [44] | 2.6B | - | 83.27 | 82.43 | 80.91 | 86.76 | 80.41 | 74.65 |
| DALL-E 3 [4] | - | - | 90.97 | 89.61 | 88.39 | 90.58 | 89.83 | 83.50 |
| SD3-Medium [14] | 2B | - | 87.90 | 91.01 | 88.83 | 80.70 | 88.68 | 84.08 |
| ***Understanding & Generation*** | | | | | | | | |
| JanusFlow [40] | 1.3B | - | 87.03 | 87.31 | 87.39 | 89.79 | 88.10 | 80.09 |
| Emu3 [72] | 8B | - | 85.21 | 86.68 | 86.84 | 90.22 | 83.15 | 80.60 |
| Janus-Pro [9] | 1.5B | 162M | 87.58 | 88.63 | 88.17 | 88.98 | 88.30 | 82.63 |
| Show-o2 [80] | 1.5B | 177M | 87.53 | 90.38 | 91.34 | 90.30 | 91.21 | **85.02** |
| Tar [20] | 1.5B | 403M | 83.59 | 89.35 | 86.91 | 93.50 | 80.80 | 82.96 |
| CHEERS | 1.5B | 83M | 90.84 | 90.24 | 89.42 | 89.71 | 87.37 | <u>83.48</u> |

generation does not rely solely on large-scale data. This highlights the effectiveness of our unified architecture, whose shared representation design enables efficient knowledge transfer between understanding and generation, resulting in robust image synthesis performance with high data efficiency.

**Progressive Improvement of Generation Capability.** To illustrate how the generation capability of CHEERS evolves throughout training, we present the progression of GenEval scores across different stages in Fig. 4. As shown in the figure, the model exhibits steady improvement as training proceeds, with clear performance gains at each stage. During the Vision-Language Alignment and General Pre-Training stages, most of the generation training data consists of real-world natural images paired with captions. Such data are complex and ambiguous, as real-world scenes often contain multiple objects, intricate interactions, and incompletely described visual details, making it difficult for the model to learn precise text-to-image correspondences. As a result, although the model gradually acquires fundamental visual semantic understanding during these stages,

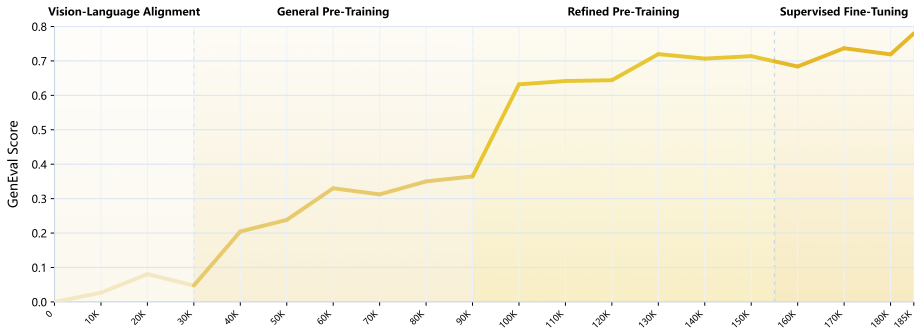


**Fig. 4:** Overall training pipeline and the progression of the GenEval score. The curve above illustrates the GenEval score as a function of cumulative training steps.

the generation performance improves moderately and remains sub-optimal until a significant surge in the Refined Pre-Training stage, where the generation training data are primarily synthetic and instruction-oriented. Compared with real-world data, synthetic data provides clearer object compositions, more explicit attribute bindings, and more direct text-image correspondence, making the learning objective better defined and easier to optimize. This significantly enhances generation fidelity and alignment, leading to rapid improvement in generation ability. Finally, during Supervised Fine-Tuning, we adopt a smaller learning rate together with a cosine decay schedule to stabilize optimization, reduce overfitting, and encourage smoother convergence. This stage further refines output quality and alignment consistency, yielding stable and incremental performance gains. Overall, these results demonstrate that the generation capability of CHEERS improves in a progressive manner, validating the effectiveness of our staged training pipeline.

### 3.3 Analysis of High-Frequency Injection

To investigate how high-frequency information contributes throughout the generation process, we visualize the high-frequency injection patterns across different denoising steps, as shown in Fig. 5 (a). The heatmaps reveal a clear temporal structure. At the early stage of generation, high-frequency components are sparsely activated and mainly concentrate around the formation of the primary object contours. As generation progresses to the intermediate stage, the magnitude of HFI slightly decreases, and the model relies primarily on semantic and low-frequency signals to complete structural details and object-level compositions. In the final stage, when the overall image structure has largely stabilized, the activation of high-frequency components increases significantly, contributing to the refinement of local textures and fine-grained visual details. This stage-wise evolution suggests that high-frequency signals are not uniformly involved, but instead dynamically modulated according to the generation stage.

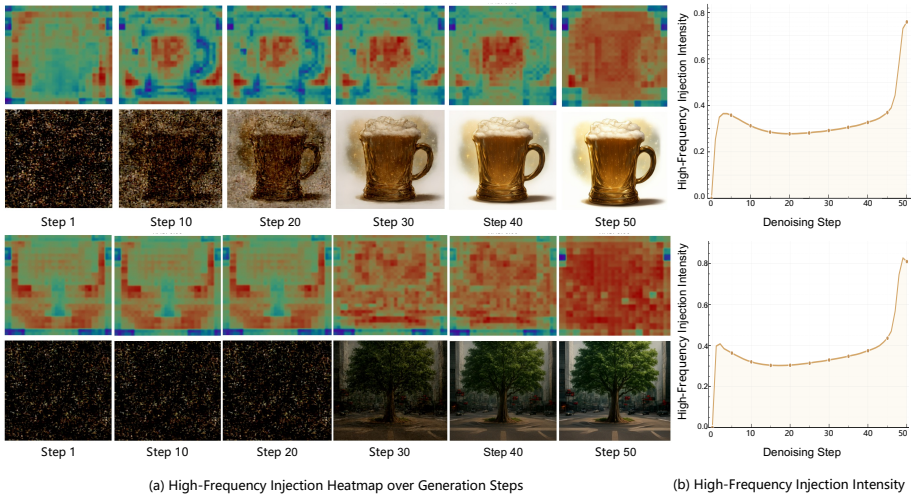


**Fig. 5:** (a) Heatmap of high frequency injection across different generation steps. (b) High frequency injection intensity at each step. The reported values are aggregated over multiple runs, with per-run normalization applied prior to averaging across samples at the same step, ensuring a faithful representation of the overall trend.

Fig. 5 (b) further quantifies the degree of high-frequency participation across the entire denoising trajectory. At the beginning of generation, the model focuses on constructing coarse layouts and global structures, during which HFI starts from a relatively low level. In the middle phase, semantic and low-frequency information guide the generation of object-level structures and attributes, resulting in a relatively moderate level of high-frequency participation. As the process approaches the final stages, the intensity and proportion of HFI increase sharply, indicating its growing role in enhancing local textures and visual fidelity. Such a progressive pattern reflects a hierarchical generation mechanism, where the model first sketches global layouts and semantic structures before gradually refining localized details and textures, resembling the coarse-to-fine process commonly observed in human drawing behavior.

### 3.4 Ablation Studies

We conduct controlled ablation experiments to investigate HFI and the impact of the generation objective on multimodal understanding. After standard Vision–Language Alignment, models are fine-tuned using 858K understanding and 850K generation samples (randomly sampled from Refined Pre-Training) for preliminary investigation and rapid validation.

**Does Generation Affect Understanding?** To further examine whether jointly training understanding and generation tasks under this architecture compromises multimodal understanding performance, we conduct a controlled experiment.

After completing the alignment stage, we fine-tune a model using only multimodal understanding data as a baseline for comparison. The results, summarized in Tab. 5, show that joint training with both understanding and generation objectives not only equips the model with image generation capability, but also achieves comparable or even slightly superior understanding performance compared to fine-tuning on understanding data alone. These findings demonstrate the effectiveness of the unified vision tokenizer design in supporting multimodal tasks within a single framework.

**Is High-Frequency Injection Necessary?** As shown in Tab. 5, we train two models under identical training settings: CHEERS and a variant without HFI. The results indicate that introducing high-frequency patch details has minimal impact on multimodal understanding performance, while leading to a substantial improvement in generation quality. Although the model without HFI is able to produce semantically consistent images, the generated results lack fine-grained visual details. In contrast, incorporating high-frequency patch details significantly enhances texture fidelity and structural sharpness. These findings suggest that HFI plays a crucial role in improving visual detail generation, making it an essential component for unified multimodal modeling.

### 3.5 Discussion

CHEERS directly leverages the pre-trained weights of a native ViT model (*e.g.*, SigLIP2), allowing the model to fully benefit from the rich representations learned during pre-training. In this way, we avoid the computational overhead of training a unified vision encoder, which has long been considered a key step in prior work for achieving unified vision-language understanding and generation. Moreover, our architecture does not freeze any core parameters, enabling all modules to be jointly fine-tuned. This ensures maximal synergy among components and allows the model to fully exploit the strengths of each module.

## 4 Related Work

### 4.1 Image Tokenizers.

UMMs require a versatile visual representation to bridge discriminative understanding and generative synthesis. The design of image tokenizers is central to this objective, revolving around the trade-offs between representation formats, feature granularities, and architectural integration.

**Table 5:** Ablation results. The first row corresponds to a model fine-tuned using understanding data only. The second and third rows report jointly trained models optimized for both understanding and generation, without and with HFI, respectively.

| Model | HFI | Fine-tuning Data | SEEDBench | MMBench | ChartQA | POPE | AI2D | GenEval | DPG-Bench |
|---|---|---|---|---|---|---|---|---|---|
| CHEERS | ✓ | Understanding | 70.8 | 65.2 | 58.5 | 87.0 | 67.7 | - | - |
| CHEERS | ✗ | Generation & Understanding | 70.0 | 66.3 | 58.8 | 86.2 | 67.3 | 0.17 | 39.11 |
| CHEERS | ✓ | Generation & Understanding | 69.8 | 67.1 | 59.9 | 87.5 | 68.1 | 0.30 | 51.63 |

**Discrete *vs.* Continuous Representations.** The choice of tokenization dictates the modeling paradigm. Discrete tokenizers like Chameleon [59] map images to a finite codebook to leverage autoregressive objectives, yet often suffer from information bottlenecks and reduced fidelity. In contrast, continuous representations in KL-regularized latent spaces preserve richer details, becoming the preference for high-fidelity generation. As highlighted in TUNA [36], continuous features, when aligned with pretrained encoders, exhibit superior efficacy for both generative and semantic understanding tasks.

**Semantic Features *vs.* High-Frequency Textures.** A fundamental disparity exists between low-frequency semantics for reasoning and high-frequency textures for reconstruction. Semantic encoders like SigLIP [65] often overlook fine-grained textures, leading to blurred synthesis, while generative VAEs lack global context. To reconcile this, Show-o [79] explores late-fusion strategies, while TUNA [36] proposes a cascaded architecture that extracts semantics directly from VAE latents to achieve a balanced and unified representation space.

**Unified visual tokenizer.** Recent efforts focus on native unification. Discrete approaches like UniTok [39] and TokLIP [31] attempt to learn shared codebooks for dual tasks but remain constrained by discretization limits. Conversely, continuous unified frameworks integrate representation encoders with generative spaces. For instance, RAE [64, 90] leverages frozen semantic features for reconstruction, while TAR [19] and SVG [49] explore joint modeling. However, achieving a seamless balance between high-level reasoning and pixel-perfect generation remains a core challenge in UMM design.

### 4.2 Unified Multimodal Models.

UMMs represent a paradigm shift toward native architectural integration, aiming to bridge the structural divide between multimodal perception and generative synthesis within a shared transformer backbone. Current research can be broadly categorized into three primary paradigms based on their modeling objectives: pure autoregressive (AR), pure diffusion, and hybrid frameworks. Pure AR models, exemplified by Chameleon [59], Emu3 [72], and Janus-Pro [9], unify modalities by quantizing visual data into discrete token sequences and optimizing via the next-token prediction objective. This approach allows them to inherit the proven scaling laws and reasoning capabilities of large language models. Conversely, pure diffusion-based UMMs such as MMaDA [81] and UniDisc [58] adopt unified masked token prediction or stochastic denoising. These models benefit from parallel decoding efficiency, which achieves significantly higher inference speeds than sequential AR, and support bidirectional reasoning for tasks like joint image-text inpainting. Finally, hybrid architectures strategically fuse these mechanisms to maximize cross-modal synergy. Models like Show-o [79], Transfusion [91], and SEED-X [16] typically maintain autoregressive modeling for sequential language while employing diffusion or flow-matching processes for continuous visual representations, thereby achieving high-fidelity generation without compromising linguistic logic.

### 4.3 Synergy in Multimodal Comprehension and Generation.

The bidirectional synergy between multimodal comprehension and generation has become a pivotal research focus. Recent studies demonstrate that discriminative understanding significantly benefits generative performance. Specifically, REPA [86] aligns diffusion transformer features with pretrained visual encoders to accelerate convergence, while VA-VAE [83] addresses the optimization dilemma between reconstruction and generation by aligning latent spaces with vision foundation models. Conversely, generative tasks increasingly serve to bolster visual comprehension. ROSS [69] introduces a reconstructive objective via latent denoising to enhance fine-grained perception and reduce hallucinations. Furthermore, UniMRG [53] incorporates auxiliary generation of intrinsic representations such as depth and segmentation to capture geometric and structural cues. These advancements suggest that internalizing generative tasks allows unified models to develop a more comprehensive understanding of spatial relations and structural layouts.

## 5 Conclusion

In this work, we present CHEERS, a novel unified multimodal model that successfully harmonizes visual comprehension and high-fidelity generation within a single framework. The core of our approach lies in the decoupling of patch-level details from semantic representations, addressing the intrinsic optimization conflict that often plagues joint multimodal modeling. By utilizing a unified vision tokenizer to extract stable semantics and a cascaded flow matching head to inject semantically gated high-frequency residuals, CHEERS ensures both robust understanding and precise image synthesis. Extensive evaluations across ten understanding benchmarks and multiple generation benchmarks demonstrate that CHEERS achieves competitive performance. Notably, our model maintains high efficiency through a 4× token compression rate and shows the emergence of zero-shot image editing capabilities, despite being trained on a relatively modest dataset of 83M samples. These results validate that the hierarchical progression from global semantic layout to localized detail refinement—resembling the human drawing process—is an effective paradigm for unified modeling. While CHEERS exhibits strong performance, future research could explore further scaling of both the LLM backbone and training data to unlock even more complex reasoning and creative generation abilities. Additionally, extending this decoupled representation framework to video understanding and generation presents a promising avenue for achieving more generalized multimodal intelligence.

**Limitation.** Despite its performance, our study has three primary constraints. First, the relatively small parameter scale of CHEERS may limit its ability to capture intricate details. Second, as it is not initialized from large-scale pre-trained VLMs, its inherent visual understanding and generation capabilities require further enhancement. Finally, the current training pipeline relies on single-image datasets; future work will incorporate more diverse and complex multimodal data to improve generalization.

## References


1. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Wu, C., Tan, H., Li, C., Yang, J., Yu, J., Wang, X., Qin, B., Wang, Y., Yan, Z., Feng, Z., Liu, Z., Li, B., Deng, J.: Llava-onevision-1.5: Fully open framework for democratized multimodal training (2025), https://arxiv.org/abs/2509.23661
2. Baade, A., Chan, E.R., Sargent, K., Chen, C., Johnson, J., Adeli, E., Fei-Fei, L.: Latent forcing: Reordering the diffusion trajectory for pixel-space image generation. arXiv preprint arXiv:2602.11401 (2026)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)
5. Chen, F., Jing, M., Lu, W., Feng, Y., Li, X., Cao, X.: Unihetero: Could generation enhance understanding for vision-language-model at large data scale? arXiv preprint arXiv:2512.23512 (2025)
6. 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)
7. 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. arXiv preprint arXiv:2506.18095 (2025)
8. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems **37**, 27056–27087 (2024)
9. 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)
10. Chu, X., Qiao, L., Zhang, X., Xu, S., Wei, F., Yang, Y., Sun, X., Hu, Y., Lin, X., Zhang, B., et al.: Mobilevlm v2: Faster and stronger baseline for vision language model. arXiv preprint arXiv:2402.03766 (2024)
11. Cui, Y., Chen, H., Deng, H., Huang, X., Li, X., Liu, J., Liu, Y., Luo, Z., Wang, J., Wang, W., et al.: Emu3. 5: Native multimodal models are world learners. arXiv preprint arXiv:2510.26583 (2025)
12. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
13. Du, S., Guo, J., Li, B., Cui, S., Xu, Z., Luo, Y., Wei, Y., Gai, K., Wang, X., Wu, K., et al.: Vqrae: Representation quantization autoencoders for multimodal understanding, generation and reconstruction. arXiv preprint arXiv:2511.23386 (2025)
14. 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)
15. Fan, W., Diao, H., Wang, Q., Lin, D., Liu, Z.: The prism hypothesis: Harmonizing semantic and pixel representations via unified autoencoding. arXiv preprint arXiv:2512.19693 (2025)

16. 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)
17. 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)
18. Gu, S., Zhang, J., Zhou, S., Yu, K., Xing, Z., Wang, L., Cao, Z., Jia, J., Zhang, Z., Wang, Y., et al.: Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. arXiv preprint arXiv:2410.18558 (2024)
19. 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)
20. 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)
21. Han, S., Kim, J.: Clip-vqdiffusion: Langauge free training of text to image generation using clip and vector quantized diffusion model. arXiv preprint arXiv:2403.14944 (2024)
22. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 (2024)
23. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European conference on computer vision. pp. 235–251. Springer (2016)
24. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
25. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2` (2025)
26. 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`
27. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)
28. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023)
29. Li, Z., Li, H., Shi, Y., Farimani, A.B., Kluger, Y., Yang, L., Wang, P.: Dual diffusion for unified image generation and understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2779–2790 (2025)
30. Liao, C., Liu, L., Wang, X., Luo, Z., Zhang, X., Zhao, W., Wu, J., Li, L., Tian, Z., Huang, W.: Mogao: An omni foundation model for interleaved multi-modal generation. arXiv preprint arXiv:2505.05472 (2025)
31. 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)
32. Lin, S., Wang, A., Yang, X.: Sdxl-lightning: Progressive adversarial diffusion distillation. arXiv preprint arXiv:2402.13929 (2024)
33. Liu, H., Yan, W., Zaharia, M., Abbeel, P.: World model on million-length video and language with blockwise ringattention. arXiv preprint arXiv:2402.08268 (2024)

34. 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)
35. Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.C., Liu, C.L., Jin, L., Bai, X.: Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences **67**(12), 220102 (2024)
36. Liu, Z., Ren, W., Liu, H., Zhou, Z., Chen, S., Qiu, H., Huang, X., An, Z., Yang, F., Patel, A., et al.: Tuna: Taming unified visual representations for native unified multimodal models. arXiv preprint arXiv:2512.02014 (2025)
37. Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al.: Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525 (2024)
38. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
39. 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)
40. Ma, Y., Liu, X., Chen, X., Liu, W., Wu, C., Wu, Z., Pan, Z., Xie, Z., Zhang, H., Yu, X., et al.: Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7739–7751 (2025)
41. Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the association for computational linguistics: ACL 2022. pp. 2263–2279 (2022)
42. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022)
43. Peng, D., Yang, X., Guo, Z., Zhang, Y., Chen, C., Zhang, Y., Yao, Y., Wan, F., Ke, W., Sun, M.: Flexivideo: Variation-aware temporal dynamics modeling for efficient video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9804–9814 (2026)
44. 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)
45. 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)
46. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
48. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8430–8439 (2019)
49. Shi, M., Wang, H., Zheng, W., Yuan, Z., Wu, X., Wang, X., Wan, P., Zhou, J., Lu, J.: Latent diffusion model without variational autoencoder. arXiv preprint arXiv:2510.15301 (2025)

50. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
51. skvarre: movie-posters-100k. https://huggingface.co/datasets/skvarre/movie_posters-100k (2023)
52. 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)
53. Su, Z., Wei, H., Cen, K., Wang, Y., Chen, G., Yuan, C., Chu, X.: Generation enhances understanding in unified multimodal models via multi-representation generation. arXiv preprint arXiv:2601.21406 (2026)
54. 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)
55. 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)
56. 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)
57. Sun, S., Zhang, Y., Song, H., Guo, Z., Chen, C., Zhang, Y., Yao, Y., Liu, Z., Sun, M.: Llava-uhd v3: Progressive visual compression for efficient native-resolution encoding in mllms. arXiv preprint arXiv:2511.21150 (2025)
58. Swerdlow, A., Prabhudesai, M., Gandhi, S., Pathak, D., Fragkiadaki, K.: Unified multimodal discrete diffusion. arXiv preprint arXiv:2503.20853 (2025)
59. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024). https://doi.org/10.48550/arXiv.2405.09818, https://github.com/facebookresearch/chameleon
60. Team, K., Bai, T., Bai, Y., Bao, Y., Cai, S., Cao, Y., Charles, Y., Che, H., Chen, C., Chen, G., et al.: Kimi k2. 5: Visual agentic intelligence. arXiv preprint arXiv:2602.02276 (2026)
61. Team, N., Han, C., Li, G., Wu, J., Sun, Q., Cai, Y., Peng, Y., Ge, Z., Zhou, D., Tang, H., et al.: Nextstep-1: Toward autoregressive image generation with continuous tokens at scale. arXiv preprint arXiv:2508.10711 (2025)
62. Team, Q.: Qwen2.5: A party of foundation models (September 2024), https://qwenlm.github.io/blog/qwen2.5/
63. 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)
64. Tong, S., Zheng, B., Wang, Z., Tang, B., Ma, N., Brown, E., Yang, J., Fergus, R., LeCun, Y., Xie, S.: Scaling text-to-image diffusion transformers with representation autoencoders. arXiv preprint arXiv:2601.16208 (2026)
65. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
66. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)

67. Wang, A.J., Mao, D., Zhang, J., Han, W., Dong, Z., Li, L., Lin, Y., Yang, Z., Qin, L., Zhang, F., et al.: Textatlas5m: A large-scale dataset for dense text image generation. arXiv preprint arXiv:2502.07870 (2025)
68. Wang, B., Lee, C., Lee, N., Lin, S.C., Dai, W., Chen, Y., Chen, Y., Yang, Z., Liu, Z., Shoeybi, M., et al.: Nemotron-cascade: Scaling cascaded reinforcement learning for general-purpose reasoning models. arXiv preprint arXiv:2512.13607 (2025)
69. Wang, H., Zheng, A., Zhao, Y., Wang, T., Ge, Z., Zhang, X., Zhang, Z.: Reconstructive visual instruction tuning. arXiv preprint arXiv:2410.09575 (2024)
70. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
71. 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)
72. 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)
73. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. arXiv:2210.14896 [cs] (2022), https://arxiv.org/abs/2210.14896
74. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. arXiv preprint arXiv:2410.13848 (2024)
75. 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. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17739–17750 (2025)
76. 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)
77. X.AI: Grok-1.5 vision preview. https://x.ai/blog/grok-1.5v (2024)
78. 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
79. 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)
80. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. arXiv preprint arXiv:2506.15564 (2025)
81. Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., Wang, M.: Mmada: Multimodal large diffusion language models. arXiv preprint arXiv:2505.15809 (2025)
82. Yao, J., Song, Y., Zhou, Y., Wang, X.: Towards scalable pre-training of visual tokenizers for generation. arXiv preprint arXiv:2512.13687 (2025)
83. 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)
84. Ye, J., Jiang, D., Wang, Z., Zhu, L., Hu, Z., Huang, Z., He, J., Yan, Z., Yu, J., Li, H., et al.: Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. arXiv preprint arXiv:2508.09987 (2025)

85. Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., Huang, F.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 13040–13051 (2024)
86. 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. arXiv preprint arXiv:2410.06940 (2024)
87. 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)
88. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
89. Zhang, J., Shen, Y., Chen, G., Song, L., Xing, E.: Dimensional collapse in vqvaes: Evidence and remedies. Advances in Neural Information Processing Systems **38**, 170934–170966 (2026)
90. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025)
91. 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)