

# GenHancer: Imperfect Generative Models are Secretly Strong Vision-Centric Enhancers

Shijie Ma<sup>1,2</sup>, Yuying Ge<sup>1,✉</sup>, Teng Wang<sup>1</sup>, Yuxin Guo<sup>1,2</sup>, Yixiao Ge<sup>1</sup>, Ying Shan<sup>1</sup>

<sup>1</sup>ARC Lab, Tencent PCG      <sup>2</sup>Institute of Automation, CAS

<https://mashijie1028.github.io/GenHancer>

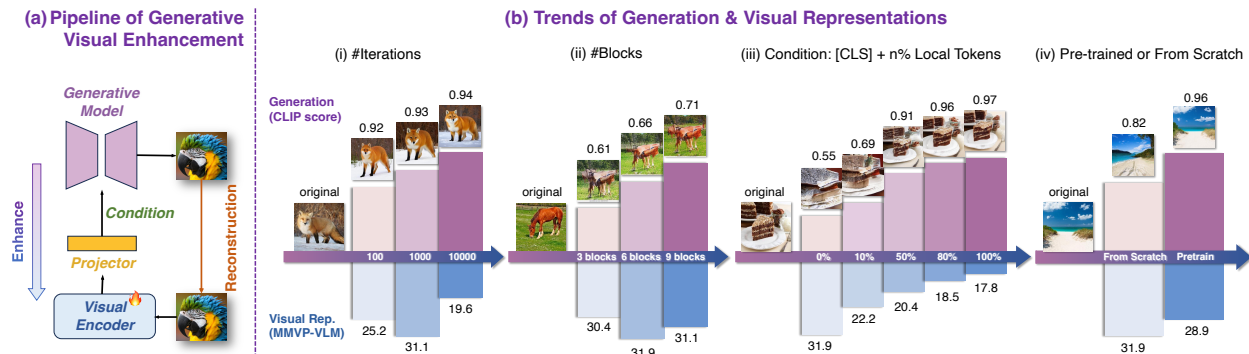


Figure 1. **Perfect generation (reconstruction) does not always yield desirable visual representations.** (a) Pipeline of fine-grained visual enhancements, where generative models take visual tokens as conditions and perform reconstruction. (b) Experiments across four dimensions, *i.e.*, training iterations, denoiser size, ratio of local tokens as conditions, and whether to use pre-trained denoisers. We measure **generation (CLIP score  $\uparrow$ )** and **visual representations (MMVP-VLM  $\uparrow$ )** performance. As the results demonstrate, although increasing the number of training iterations, adding more denoiser blocks, using a larger ratio of local tokens as conditions, and employing pre-trained denoisers lead to better generation results, the performance of visual representations does not always improve. Best viewed zoomed in.

## Abstract

The synergy between generative and discriminative models receives growing attention. While discriminative Contrastive Language-Image Pre-Training (CLIP) excels in high-level semantics, it struggles with perceiving fine-grained visual details. Generally, to enhance representations, generative models take CLIP’s visual features as conditions for reconstruction. However, the underlying principle remains underexplored. In this work, we empirically found that **visually** perfect generations are not always optimal for representation enhancement. The essence lies in effectively extracting fine-grained knowledge from generative models while mitigating irrelevant information. To explore critical factors, we delve into three aspects: (1) *Conditioning mechanisms*: We found that even a small number of local tokens can drastically reduce the difficulty of reconstruction, leading to collapsed training. We thus conclude that utilizing **only** global visual tokens as conditions is the most effective strategy. (2) *Denoising configurations*: We observed that end-to-end training introduces extraneous information. To address this, we propose a two-stage training strategy to prioritize learning useful visual knowledge.

Additionally, we demonstrate that lightweight denoisers can yield remarkable improvements. (3) *Generation paradigms*: We explore both continuous and discrete denoisers with desirable outcomes, validating the versatility of our method. Through our in-depth explorations, we have finally arrived at an effective method, namely GenHancer, which consistently outperforms prior arts on the MMVP-VLM benchmark, *e.g.*, 6.0% on OpenAI CLIP. The enhanced CLIP can be further plugged into multimodal large language models for better vision-centric performance. All the models and codes are made publicly available.

## 1. Introduction

Generative and discriminative models have evolved rapidly in recent years [3, 30, 50, 62]. Both of them exhibit complementary strengths, where generative models like diffusion models [16, 42, 63] and rectified flow [9, 29] capture low-level visual details, while discriminative models like Contrastive Language-Image Pre-Training (CLIP) [41, 66] and DINO [39] excel in high-level semantics. This complementary nature enables a synergistic relationship between them. Pioneering work [65] has shown that discriminative mod-

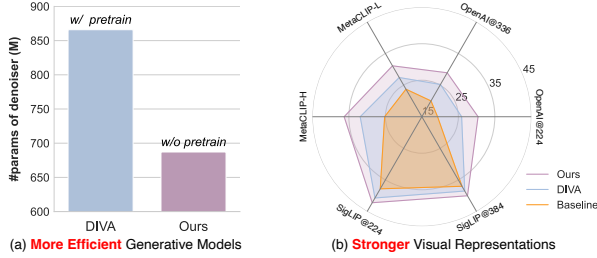


Figure 2. Comparison with prior method [58]. (a) We only need a lightweight denoiser, but (b) achieve stronger performance than DIVA [58], which relies on pre-trained heavy generative models.

els can facilitate the training of generative models through feature alignment. Conversely, generative models can also enhance discriminative models by improving their ability to understand fine-grained patterns, *e.g.*, orientation, color and quantity. This enhancement is particularly pertinent for models like CLIP, which have inherent visual shortcomings [53] that could also limit Multimodal Large Language Models (MLLMs) [31, 52] in vision-centric tasks. Recent works [18, 57, 58] have attempted to enhance CLIP ViT by using the visual features of ViT [7] as conditional inputs for generative models. These models perform self-supervised reconstruction to compel the discriminative model to capture fine-grained visual details, as in Fig. 1 (a). While they demonstrate the potential of enhancing representations with generative models, they often rely on pre-trained heavy denoisers and do not explore the underlying principle.

To enable generative models to enhance visual representations, a natural question arises: *Do we need a perfect generative model to achieve this enhancement?* To address this question, we conducted preliminary experiments from several dimensions, including training iterations, local tokens ratio as conditions, denoiser sizes, and whether to use a pre-trained generative model, as in Fig. 1 (b). The answer is that perfect generation (reconstruction) does not always yield desirable representations. As in Fig. 1 (iii), introducing more local tokens as conditions can improve reconstruction, while the visual enhancement will be drastically degraded. In Fig. 1 (iv), although the pre-trained denoiser exhibits better reconstruction, its representations are weaker.

This leads us to further investigate the key points for generative models to effectively enhance visual representations. We argue that generative models simultaneously contain useful knowledge, like visual patterns and details, as well as irrelevant information, like the gap between CLIP’s feature space and generative models’ condition space. To effectively enhance representations, our general *philosophy* is that discriminative models should prioritize learning useful knowledge from generative models while circumventing irrelevant information. Furthermore, generative models can be divided into continuous [29, 32] and discrete [8] ones, with different denoising objectives, which should also

be considered. Consequently, we conduct in-depth explorations from three key aspects: conditioning mechanisms, denoising configurations, and generation paradigms.

**Key Point #1: Which part of the visual information should generative models focus on?** As in Fig. 1 (a), generative models take visual tokens as conditions. The choice of different tokens significantly impacts the outcomes. In this regard, we find *only* the global token (*i.e.*, class token) could yield desirable enhancements. We attribute this to the fact that class token *alone* helps maximize mutual information between CLIP ViT and generative models, while local tokens bring about information leakage and drastically reduce the task’s difficulty, resulting in collapsed learning.

**Key Point #2: How to design denoising configurations to transfer useful information for visual representations?** The structure of the denoiser could determine the enhancement effects. Additionally, it is essential to mitigate irrelevant information. Therefore, we investigate the influence of different denoiser sizes and training stages. In this paper, we propose GenHancer, a two-stage post-training method for visual enhancements. In the first stage, we pre-train the projector and denoiser while freezing the ViT, learning basic reconstruction abilities and mitigating irrelevant information. In the second stage, we fine-tune CLIP ViT to enhance its fine-grained visual representations. Meanwhile, we empirically found that a lightweight denoiser is sufficient to achieve remarkable results, which is more efficient yet stronger, as in Fig. 2.

**Key Point #3: Do two types of denoisers share a common enhancing principle for visual representations?** For both continuous and discrete denoisers, we present tailor-made designs, including denoiser and conditioning structure. Moreover, we reveal that previous Key Points #1, #2 apply to both types, indicating the versatility of our method.

Our contributions are summarized as follows:

- We conduct an in-depth study on visual representation enhancements with generative models and make the innovative discovery that perfect reconstruction and pre-trained models are not necessary. This leads us to explore three key aspects: conditioning mechanisms, denoising configurations, and the generation paradigms.
- We propose GenHancer, a two-stage post-training method with only lightweight denoisers for visual enhancements, which uses only the class token as the conditional input to perform self-supervised reconstruction. Our method is applicable to both continuous and discrete denoisers.
- Comprehensive vision-centric evaluations show that our enhanced CLIP significantly outperforms prior methods that rely on pre-trained heavy denoisers, as in Fig. 2.

## 2. Related Works

**MLLMs and Vision Encoders.** Currently, MLLMs [15, 28] predominantly employ CLIP [41, 47] for visual encod-

ing. Tong *et al.* [53] identified several failure patterns in CLIP, which hinder the fine-grained visual understanding. To overcome this issue, early efforts [20, 52, 53] employed an ensemble of visual experts to combat the visual shortcomings. More recently, ROSS [57] leverages intrinsic visual activations and incorporates a self-supervised visual reconstruction loss during training MLLMs. Complementarily, DIVA [58] proposes to enhance CLIP’s fine-grained abilities through diffusion feedback. Similar to [58], we independently enhance CLIP’s internal representations, which not only strengthen CLIP as a vision-language retriever but also enable the enhanced CLIP to be seamlessly integrated into MLLMs in a *plug-and-play* manner for better fine-grained vision-centric performance.

**Enhancing Visual Representations with Diffusion Models.** Early works [34, 44, 51] utilize generative models as data augmenters [36, 37, 49]. Another line of works [5, 18, 59] leverages self-supervised reconstruction tasks with diffusion models, which helps models grasp visual details and learn fine-grained representations. Similarly, DIVA [58] takes CLIP’s features as conditional inputs to the diffusion model [42], addressing its visual shortcomings through reconstruction. In summary, prior arts predominantly rely on diffusion models [11, 42], whereas we apply our method to both continuous and discrete generative models.

**Vision-Centric Benchmarks.** Canonical evaluations of MLLMs focus on fundamental multimodal Q&A capabilities across various domains, *e.g.*, general perception and cognition [10], text and characters [45], scientific fields [33], and potential hallucinations [13, 27] in MLLMs. However, these benchmarks could not effectively assess a model’s fine-grained [14, 15] visual perception abilities, such as object color, quantity, orientation, and viewpoint. To solve this issue, Tong *et al.* [53] systematically explore the failure modes of CLIP and propose a challenging MMVP benchmark with 9 visual patterns. CV-Bench [52] further expands with 2,600 vision-centric VQA questions, covering dimensions like spatial relationships, count, depth, and distance of both 2D and 3D domains. Besides, NaturalBench [26] curates natural adversarial samples that are easy for humans but MLLMs struggle with. In this paper, we employ these vision-centric benchmarks to comprehensively evaluate models’ fine-grained visual abilities.

### 3. Preliminaries of Generative Models

In principle, generative models can be divided into continuous and discrete ones. For continuous generative models, we focus on the recently popular rectified flow [29, 32], while discrete generative models are conventionally built upon pre-trained codebooks [8, 55] for discrete modeling.

**Rectified Flow (RF).** Most generative models explicitly or implicitly learn a mapping from a basic distribution,

*e.g.*, Gaussian distribution  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ , to a target distribution, typically the real data distribution  $p_{\text{data}}$ . The core idea of RF is to learn an Ordinary Differential Equation (ODE)  $dZ_t = \mathbf{u}(Z_t, t)dt$  that follows a straight path from  $\pi_0$  to  $\pi_1$ . Here  $\mathbf{u}(Z_t, t)$  is a time-conditional velocity field. This could be achieved by solving a least squares regression problem:  $\min_{\mathbf{u}} \int_0^1 \mathbb{E} [\| (X_1 - X_0) - \mathbf{u}(X_t, t) \|^2] dt$ , where  $X_t = tX_1 + (1-t)X_0$ . In practice, we use  $\phi$  to parameterize  $\mathbf{u}$ , and  $t$  is basically sampled from the uniform distribution  $\mathcal{U}(0, 1)$ . The learning objective of RF is:

$$\mathcal{L}_{\text{RF}} = \mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_1} \left\| (\mathbf{x}_1 - \mathbf{x}_0) - \mathbf{u}_{\phi}(t\mathbf{x}_1 + (1-t)\mathbf{x}_0, t) \right\|_2^2, \quad (1)$$

where  $t \sim \mathcal{U}(0, 1)$ ,  $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ,  $\mathbf{x}_1 \sim p_{\text{data}}$ .

**Discrete Generative Models.** For discrete modeling, one should first learn a discrete codebook, where images are represented by their corresponding indices. For example, VQ-GAN [8] employs some schemes [12, 67] to learn a discrete codebook of perceptually rich representations. Subsequently, given indices  $s_{<i}$  of image  $\mathbf{x}$ , the discrete generative model  $\mathbf{p}_{\phi}$  learns to predict the categorical distribution of the next index  $s_i$  via the cross-entropy objective:

$$\mathcal{L}_{\text{CE}} = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} - \log \prod_{i=1}^L \mathbf{p}_{\phi}(s_i | s_{<i}), \quad (2)$$

where  $L$  denotes the sequence length of a sample.  $\mathbf{p}_{\phi}$  could be any form of model capable of modeling discrete distributions, *e.g.*, PixelCNNs [54] and Transformers [8, 56].

**Conditional Generation.** To achieve conditional generation, one could incorporate the condition  $\mathbf{c}$ , *e.g.*, class labels or text prompts, into the parameterized model in Eq. (1) and Eq. (2) as  $\mathbf{u}_{\phi}(\mathbf{x}_t, t, \mathbf{c})$  and  $\mathbf{p}_{\phi}(s_i | s_{<i}, \mathbf{c})$ , respectively.

## 4. Method

### 4.1. Overview and Formulation

**Overview.** We propose a two-stage post-training method, namely GenHancer, to enhance CLIP ViT’s fine-grained representations, as in Fig. 3 (a). To capture key information from generative models, we delve into three aspects: First, the choice of visual tokens for condition determines the difficulty of the reconstruction task, which is crucial for enhancement (Sec. 4.2). Second, we introduce denoising configurations, which enable ViT to capture useful knowledge while mitigating irrelevant information (Sec. 4.3). Third, we present tailored design for both continuous and discrete generative models (Sec. 4.4), also shown in Fig. 3 (b), (c).

**Notations.** Here, two types of generative models are uniformly represented as  $\mathbf{g}_{\phi}$  parameterized by  $\phi$ . Let  $\mathbf{v}_{\theta}$  denote CLIP’s visual encoder with parameters  $\theta$ , whose features are connected to  $\mathbf{g}_{\phi}$  as conditions through projector  $\mathbf{h}_{\omega}$ , *i.e.*,  $\mathbf{h}_{\omega} \circ \mathbf{v}_{\theta}(\mathbf{x})$ . The input sample is  $\mathbf{x}$ , which becomes

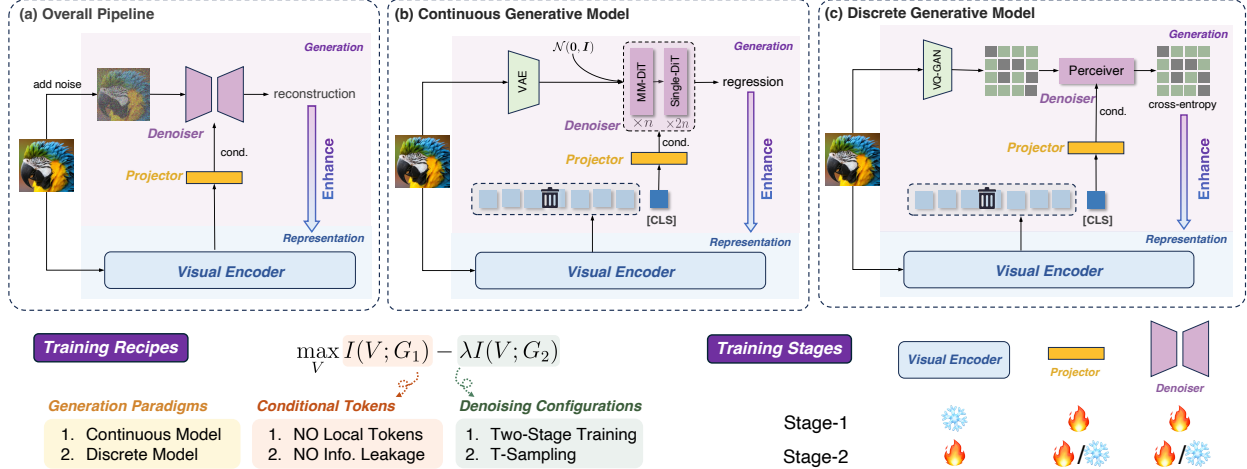


Figure 3. The two-stage post-training framework for visual enhancements. (a) Overall training pipeline. (b) Continuous generative model as the denoiser. We employ a lightweight FLUX-like DiT [24] (but with fewer blocks) and employ a regression loss of flow matching. (c) Discrete generative model as the denoiser. We choose a lightweight Perceiver [19] and employ cross-entropy loss to predict masked tokens.

$\tilde{x}$  in the denoising space, *e.g.*, VAE [21] and VQ-GAN [8] for continuous and discrete denoisers, respectively.

**Repurposing Conditional Generation to Self-supervised Reconstruction.** Generative models can capture low-level details. To transfer this capability to  $v_\theta$ , we replace the original condition  $c$  with the visual feature  $v_\theta(x)$ . By reconstructing the visual inputs,  $v_\theta$  learns to grasp low-level visual details and is enhanced with fine-grained representations. In this sense, we transform the original conditional generation into a self-supervised reconstruction task. The learning objectives for continuous  $\mathcal{L}_c$  and discrete generative models  $\mathcal{L}_d$  can be re-written as Eq. (3) and Eq. (4):

$$\mathcal{L}_c = \mathbb{E}_{t, \mathbf{x}, \tilde{\mathbf{x}}_0, \tilde{\mathbf{x}}_1} \left\| (\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_0) - \mathbf{g}_\phi(\tilde{\mathbf{x}}_t, t, \mathbf{h}_\omega \circ \mathbf{v}_\theta(\mathbf{x})) \right\|_2^2, \quad (3)$$

where  $t \sim \mathcal{U}(0, 1)$ ,  $\tilde{\mathbf{x}}_t = t\tilde{\mathbf{x}}_1 + (1-t)\tilde{\mathbf{x}}_0$ ,

$$\mathcal{L}_d = \mathbb{E}_{\mathbf{x}} - \log \prod_{i=1}^L \mathbf{g}_\phi(s_i | s_{<i}, \mathbf{h}_\omega \circ \mathbf{v}_\theta(\mathbf{x})). \quad (4)$$

Here,  $\mathbf{h}_\omega \circ \mathbf{v}_\theta(\mathbf{x})$  serves as the conditional input of  $\mathbf{g}_\phi$ .

**Formulation.** Let  $G$  and  $V$  denote random variables of features of  $\mathbf{g}_\phi$  and  $\mathbf{v}_\theta$ .  $I(\cdot)$  and  $H(\cdot)$  denote mutual information and entropy. Then we have the following theorem:

**Theorem 1.** When  $\mathbf{g}_\phi$  is fixed, self-supervised reconstruction is equivalent to maximizing the mutual information  $I(V; G)$  between  $V$  and  $G$ . The knowledge learned by  $\mathbf{v}_\theta$  from  $\mathbf{g}_\phi$  can be interpreted as the increase in  $I(V; G)$ .

*Proof.* The mutual information could be written as:  $I(V; G) = H(G) - H(G|V)$ . Through reconstruction in Eq. (3) or Eq. (4), by conditioning  $G$  on  $V$ ,  $V$  is trained to approximate the distribution of  $G$ . Consequently,  $H(G|V)$  decreases during training. While  $H(G)$  is fixed, the decrease in  $H(G|V)$  leads to the increase in  $I(V; G)$ .  $\square$

From the results in Fig. 1 (b)(i), the reconstruction improves as training progresses, which corresponds to an increasing  $I(V; G)$ . However, visual representations might decrease. In light of this, for the enhancement of visual representations, the knowledge in  $G$  can be decomposed into useful knowledge  $G_1$  (*e.g.*, basic semantics, visual patterns) and irrelevant information  $G_2$  like the gap between feature space of  $\mathbf{v}_\theta$  and condition space of  $\mathbf{g}_\phi$ . In this regard, to effectively enhance visual representations, **our underlying philosophy is: The visual encoder should learn to capture useful knowledge from generative models as much as possible, *i.e.*,  $\max I(V; G_1)$ , while avoiding irrelevant information, *i.e.*,  $\min I(V; G_2)$ .** This equals to applying regularization on  $V$  to prevent overfitting to  $G_2$ :

$$\max_V I(V; G_1) - \lambda I(V; G_2) \Rightarrow \max_V I(V; G_1) + \lambda d(V; V_0), \quad (5)$$

$V_0$  is the initial visual model and  $d(\cdot)$  is a distance metric.

## 4.2. Conditional Visual Tokens

The choice of conditional visual tokens is crucial for visual enhancement. If too many tokens are fed to the generative model, the reconstruction becomes excessively easy. The reason is that local tokens directly correspond to image areas with information leakage. In this case,  $I(V; G_1)$  in Eq. (5) becomes small and  $\mathbf{v}_\theta$  fails to grasp useful information from  $\mathbf{g}_\phi$ . To ensure a remarkable  $I(V; G_1)$ , we argue that the number of local tokens should be carefully controlled. Our experiments show that even a small number of local tokens, though achieving good reconstruction quality, can still cause marginal visual enhancement, as in Fig. 1 (iii). As a result, we propose that the visual condition features should exclusively comprise *only* the class token [CLS]. This strategy applies to both continuous and discrete models, as validated in Fig. 5 of Sec. 5.3.

### 4.3. Denoising Configurations

To effectively enhance visual representations, we aim to maximize  $I(V; G_1)$  while suppressing  $I(V; G_2)$  in Eq. (5). In this regard, our explorations are three-fold: training stages, timestamp sampling of the continuous denoiser, and the update strategy for  $v_\theta$ .

**Two Stage Training.** An important source of  $G_2$  is the gap between the feature space of  $v_\theta$  and the conditions of  $g_\phi$ , which is irrelevant to representation learning and could degrade the performance. Furthermore, since  $g_\phi$  is lightweight and randomly initialized, it could introduce potential noise to  $v_\theta$  at the beginning. Consequently, we propose a two-stage training pipeline. At Stage-1, we train the denoiser  $g_\phi$  and the projector  $h_\omega$  while freezing  $v_\theta$ , in which  $g_\phi$  acquires basic generative capabilities for visual enhancements and  $h_\omega$  learns to bridge the space gap, thereby reducing  $I(V; G_2)$ . In Stage-2, we focus on enlarging  $I(V; G_1)$  and train  $v_\theta$  to improve fine-grained representations. Moreover, we empirically found that as long as Stage-1 is performed sufficiently, the impact of whether the denoiser and projector are trained in Stage-2 is negligible.

**Low Rank Adaption (LoRA) of  $v_\theta$ .** The pre-trained visual encoder  $v_\theta$  possesses strong global semantics, *i.e.*,  $V_0$ , which should be maintained when incorporating fine-grained perception. To prevent  $v_\theta$  from overfitting during reconstruction, we update  $v_\theta$  using LoRA [17], which implicitly constrains  $d(V, V_0)$  in Eq. (5).

**Timestamp Sampling.** For continuous models like RF, timestamp sampling is of vital importance. Conventionally, RF [32] is trained to predict velocity across timestamps uniformly in  $[0, 1]$ . Considering  $x_t = tx_1 + (1-t)x_0$ , prior works [9] uncover that the velocity target at intermediate timestamps, *i.e.*,  $t \approx 0.5$ , is more challenging. In our case, sampling intermediate timestamps more frequently could increase the difficulty of the reconstruction task, thus effectively amplifying  $I(V; G_1)$  and allowing the visual encoder  $v_\theta$  to effectively acquire useful fine-grained knowledge from  $G_1$ . In this regard, we propose *scaled* Logit-Normal sampling for timestamps, as shown below:

$$t = \text{sigmoid}(s \cdot \varepsilon), \quad \text{where } \varepsilon \sim \mathcal{N}(0, 1). \quad (6)$$

Here,  $\varepsilon$  is sampled from the normal distribution,  $\text{sigmoid}(x) = \frac{1}{1 + \exp(-x)}$ , and  $s > 0$  is the scale hyperparameter that controls the extent to which sampling is focused on the intermediate timestamps. Smaller  $s$  results in more frequent sampling around 0.5. The diagrams of distributions in various  $s$  are illustrated in the Appendix.

### 4.4. Generation Paradigms

For both types of generative models, we need to design architectures for denoisers and the implementation of the conditioning mechanism. Notably, our denoiser is lightweight

and randomly initialized, without pre-trained weights of heavy denoisers like Stable Diffusion [42] in [58].

**Continuous Generative Models.** We choose RF as the continuous denoiser, which is modeled in the latent space of pre-trained VAE [21]. The structure is inherited from FLUX.1-dev [24], consisting of  $n \times$  Multimodal Diffusion Transformer (MM-DiT) [9, 40] blocks and  $2n \times$  single-stream DiT (Single-DiT) blocks, as shown in Fig. 3 (b). By default, we set  $n = 2$ , which is very efficient with  $\sim 1/10$  parameters of the original FLUX.1-dev denoiser. Similar to DiT [9, 40], the condition of visual tokens ( $[\text{CLS}]$  of  $v_\theta$ ) is introduced through the modulation mechanism via adaptive layernorm [1, 40]. The learning objective is the regression of flow matching in Eq. (3).

**Discrete Generative Models.** Here, we choose Perceiver [19] as the discrete denoiser, building upon off-the-shelf VQ-GAN’s codebook [8]. We first mask a certain proportion of input tokens. The condition of visual features is introduced via a cross-attention module, as depicted in Fig. 3 (c). Specifically, we set the query as the unmasked tokens  $s_{<i}$ , while the key and value are the concatenation of the unmasked tokens and  $[\text{CLS}]$  of  $v_\theta$ . They are collectively fed to the Perceiver with cross-entropy loss to predict the masked token indices  $s_i$ , as in Eq. (4).

## 5. Experiments

### 5.1. Experimental Setup

**Implementation Details.** For continuous generative models, we choose RF, whose structure is similar to FLUX.1-dev [24], but with only 2 MM-DiT and 4 Single-DiT blocks ( $\sim 10\%$  of the parameters). The discrete denoiser is parameterized by a 6-layer Perceiver to predict the masked tokens indexed by VQ-GAN’s codebook [8]. Similar to [61], the mask ratio is randomly sampled from 50% to 90%. For both generative models, we only take the  $[\text{CLS}]$  token of CLIP ViT as the conditional input while dropping other local tokens to prevent information leakage. We choose the scale factor in Eq. (6) as 1 by default.

**Training Details.** Our training process consists of two stages, each involving one epoch on the CC3M [43] dataset. We choose AdamW as the optimizer, with a learning rate of  $1e-4$  and  $1e-5$  for Stage-1 and Stage-2, respectively. At Stage-2, we optimize the visual encoder using LoRA with a rank of 16. We employ a global batch size of 256.

**Comparative Baseline.** Similar to [58], our method GenHancer independently enhances CLIP via post-tuning. When equipped with our enhanced CLIP and trained with original recipes, MLLMs could perform better on vision-centric benchmarks. In this regard, GenHancer could be viewed as a *plug-and-play* vision-enhancement method for MLLMs. We primarily compare with DIVA [58].

Table 1. Performance of various CLIP backbones in MMVP-VLM benchmark. Here, we report our results using the continuous denoiser. The enhanced CLIP consistently outperforms prior methods across various visual patterns. The visual patterns are symbolized as: 🕒: Orientation and Direction, 🔍: Presence of Specific Features, 🔄: State and Condition, 📊: Quantity and Count, 📍: Positional and Relational Context, 🎨: Color and Appearance, ⚙️: Structural and Physical Characteristics, 📄: Texts, 📷: Viewpoint and Perspective.

CLIP Backbone	#Params (M)	Resolution	Method	🕒	🔍	🔄	📊	📍	🎨	⚙️	📄	📷	Average
OpenAI ViT-L-14	427.6	224 <sup>2</sup>	Original	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
			+ DIVA	13.3	20.0	40.0	6.7	20.0	53.3	46.7	20.0	13.3	25.9
			+ Ours	13.3	33.3	33.3	20.0	6.7	73.3	46.7	20.0	40.0	<b>31.9 (+6.0)</b>
OpenAI ViT-L-14	427.9	336 <sup>2</sup>	Original	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
			+ DIVA	26.7	20.0	33.3	13.3	13.3	46.7	26.7	6.7	40.0	25.2
			+ Ours	6.7	20.0	33.3	20.0	6.7	73.3	53.3	26.7	26.7	<b>29.6 (+4.4)</b>
MetaCLIP ViT-L-14	427.6	224 <sup>2</sup>	Original	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
			+ DIVA	6.7	6.7	60.0	0.0	26.7	66.7	20.0	20.0	40.0	27.4
			+ Ours	13.3	20.0	53.3	13.3	26.7	80.0	33.3	13.3	33.3	<b>31.9 (+4.5)</b>
MetaCLIP ViT-H-14	986.1	224 <sup>2</sup>	Original	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
			+ DIVA	13.3	20.0	53.3	33.3	13.3	66.7	33.3	13.3	40.0	31.9
			+ Ours	20.0	20.0	66.7	26.7	26.7	66.7	33.3	20.0	53.3	<b>37.0 (+5.1)</b>
SigLIP ViT-SO-14	877.4	224 <sup>2</sup>	Original	26.7	20.0	53.3	40.0	20.0	66.7	40.0	20.0	53.3	37.8
			+ DIVA	13.3	26.7	60.0	46.7	13.3	73.3	53.3	26.7	53.3	40.7
			+ Ours	20.0	20.0	66.7	60.0	20.0	86.7	40.0	13.0	53.3	<b>42.2 (+1.5)</b>
SigLIP ViT-SO-14	878.0	384 <sup>2</sup>	Original	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
			+ DIVA	26.7	33.3	53.3	26.7	13.3	80.0	40.0	26.7	46.7	38.5
			+ Ours	26.7	20.0	66.7	33.3	13.3	86.7	40.0	26.7	46.7	<b>40.0 (+1.5)</b>

**Evaluation Protocol.** Following [58], we perform visual enhancements on six CLIP backbones, including OpenAI CLIP ViT-L @224/@336 [41], MetaCLIP@224 ViT-L/H [62] and SigLIP-SO-14 @224/@384 [66]. We use MMVP-VLM [53] to evaluate fine-grained perception abilities. Subsequently, we follow the official training recipes of LLaVA-1.5 [31] to train MLLMs with our enhanced CLIP ViT. The resulting MLLMs are comprehensively evaluated on vision-centric benchmarks like MMVP-MLLM [53], CV-Bench [52] and NaturalBench [26], as well as multimodal understanding benchmarks, including POPE [27] ScienceQA [33] and HallusionBench [13].

## 5.2. Comparative Results

**Our method significantly enhances CLIP’s fine-grained visual perception abilities.** We evaluate CLIP models on the challenging MMVP-VLM benchmark [53], which contains 9 fine-grained visual patterns for a comprehensive vision-centric evaluation. As in Table 1, our method with only a lightweight denoiser surpasses the previous method [58] that employed a heavy pre-trained denoiser across multiple CLIP backbones, with variations in resolution and parameters. For example, our method outperforms DIVA by 6.0% and 4.5% on OpenAI CLIP and MetaCLIP, respectively. Besides, CLIP’s visual shortcomings are effectively addressed after post-training, *e.g.*, we improved MetaCLIP’s color perception (🎨) from 46.7% to 80.0%, and enhanced its viewpoint understanding (📷) by 20%.

**Qualitative Evaluations.** We present two cases in Fig. 4. Although DIVA achieves better reconstructions, our method

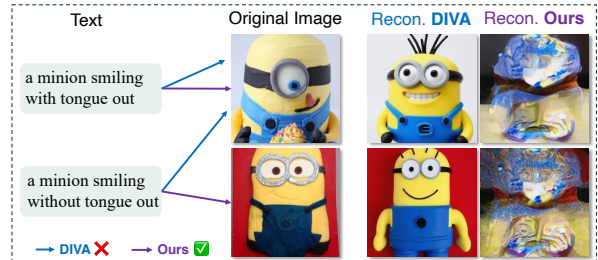


Figure 4. Qualitative results. Although DIVA achieves better reconstructions of input images, it fails to perceive fine-grained visual details between ‘tongue out’ and ‘without tongue out’.

correctly retrieves images for given texts, while DIVA fails. This further emphasizes that better reconstruction does not necessarily lead to better representations.

### Plug-and-play vision-centric enhancements for MLLMs.

Our method *independently* enhances CLIP ViT with fine-grained representations. Considering that existing MLLMs [2, 30, 31] predominantly use CLIP ViT as the visual encoder, we replace the original CLIP with the enhanced CLIP as a *plug-and-play* module and integrate it into MLLMs to explore the impact of the enhanced visual representations on MLLMs’ final performance. For fair comparisons, we adopt the same training setup as LLaVA-1.5 [31], *i.e.*, training data and stages, to train MLLMs. For DIVA [58], we adopt the official CLIP checkpoints. We conduct a comprehensive evaluation of the MLLMs on multiple vision-centric benchmarks, including MMVP-MLLM [53], CV-Bench [52] and NaturalBench [26], as well as some general multimodal understanding bench-

Table 2. Comprehensive evaluation of MLLMs (LLaVA-1.5 [31]), including vision-centric and conventional MLLM benchmarks. † We use official DIVA CLIP checkpoints [58] to reproduce the results. ‡ Similar to [25], we select the choice with the highest likelihood as MLLM’s prediction. Hallusion: HallusionBench [13]. SciQA: ScienceQA [33]. **Bold** and underline indicate the best and the second best.

LLM	CLIP	Vision-Centric Benchmarks							Conventional MLLM Benchmarks					
		MMVP-MLLM [53]	NaturalBench [26]‡				CV-Bench 2D [52]		CV-Bench 3D [52]	POPE [27]			SciQA-IMG [33]	Hallusion Avg. [13]
			Acc	Q-Acc	I-Acc	G-Acc	ADE20K	COCO		rand	pop	adv		
Vicuna-7B	Original	24.7	<u>76.4</u>	<u>53.6</u>	<u>56.4</u>	17.6	49.6	60.9	58.7	87.3	86.1	84.2	<b>66.8</b>	27.6
	DIVA†	<b>31.3</b>	75.3	51.7	56.1	22.3	51.3	63.4	<u>60.2</u>	<u>87.9</u>	<b>87.0</b>	<b>84.6</b>	66.3	<b>28.6</b>
	Ours	<u>30.7</u>	<b>77.3</b>	<b>55.6</b>	<b>59.1</b>	<b>24.4</b>	<b>52.9</b>	<b>63.6</b>	<b>63.2</b>	<b>88.1</b>	<u>86.7</u>	<b>84.6</b>	<u>66.5</u>	<u>28.4</u>
Vicuna-13B	Original	30.7	<u>76.3</u>	<u>52.9</u>	55.1	13.8	52.6	63.3	65.0	87.1	86.2	84.5	71.6	24.5
	DIVA†	35.3	76.0	52.7	<u>56.0</u>	16.8	53.2	<b>64.3</b>	<u>65.8</u>	<b>88.1</b>	<b>87.4</b>	84.8	71.8	<u>25.2</u>
	Ours	<b>36.7</b>	<b>77.2</b>	<b>55.3</b>	<b>58.7</b>	<b>22.9</b>	<b>55.3</b>	<b>64.3</b>	<b>66.4</b>	<u>87.8</u>	<u>87.0</u>	<b>84.9</b>	<b>72.3</b>	<b>26.4</b>

Table 3. Performance of zero-shot classification and retrieval that require global semantics. We report the results of original and post-tuned OpenAICLIP@224.

Method	Classification			Retrieval-Image@5		Retrieval-Text@5		
	IN-1K	C100	SUN397	Cars	Flickr30k	COCO	Flickr30k	COCO
Original	75.5	76.1	67.5	77.7	87.2	61.1	97.4	79.2
Ours	75.6	76.1	67.5	77.6	87.3	61.2	97.2	79.4

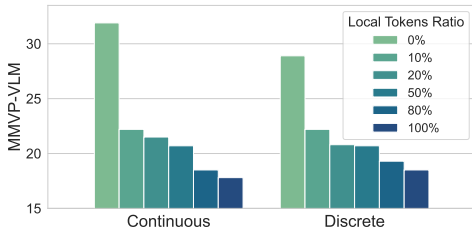


Figure 5. Performance of CLIP across various conditional visual tokens on MMVP-VLM, *i.e.*,  $[\text{CLS}] + n\% [\text{LOCAL}]$ .

marks. Results in Table 2 show that visual enhancement of CLIP is effectively transferred to MLLMs, resulting in significant improvements across vision-centric benchmarks. For instance, compared to the original CLIP in Vicuna-7B MLLM, we achieved 6.0% and 4.5% improvements on MMVP-MLLM and CV-Bench 3D, respectively.

**Visual enhancements do not hurt CLIP’s original global semantics.** CLIP has inherently strong global semantics in classification-based tasks [35, 38]. To explore how fine-grained enhancements affect this ability, we evaluate zero-shot classification on datasets like ImageNet-1K [6], CIFAR100 [23], Stanford Cars [22], and SUN397 [60] and zero-shot cross-modal retrieval tasks on Flickr30k [64] and COCO [4]. Table 3 reveals that the performance difference is minimal ( $< 0.3\%$ ) across various settings, which means that our method could enhance CLIP’s fine-grained understanding without forgetting its global semantics [46, 48].

### 5.3. Key Explorations and Ablations

**Key Point #1: Selecting Conditional Visual Tokens.** As in Sec. 4.2, selecting conditional visual tokens is critical for enhancing representations. We conduct experiments by choosing the class token and different proportions of

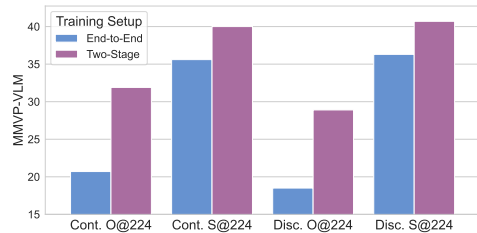


Figure 6. Comparison of CLIP with end-to-end and the proposed two-stage training on MMVP-VLM. Here, Cont. and Disc. denote continuous and discrete denoisers. O: OpenAICLIP. S: SigLIP.

local tokens, *i.e.*,  $[\text{CLS}] + n\% [\text{LOCAL}]$ . As displayed in Fig. 5, even a very small ratio (10%) leads to significant performance degradation, which suggests that local tokens carry substantial signals for reconstruction, making the task too easy with information leakage. Consequently, this prevents the visual encoder from effectively learning fine-grained details and brings about a limited  $I(V; G_1)$ . The conclusion applies to both types of  $g_\phi$ . Therefore, we propose to choose only the class token as the condition.

**Key Point #2.1: Two-Stage Training.** As elaborated in Sec. 4.3, in Stage-1 of the two-stage training scheme, the projector learns to bridge the gap between the feature space of the visual encoder and the condition space of the denoiser, which serves as irrelevant information  $G_2$ . Ablations comparing end-to-end with the proposed two-stage training are illustrated in Fig. 6. End-to-end training consistently exhibits a performance drop of over 5% across various settings. This indicates that our two-stage training is crucial in preventing the interference from  $G_2$ .

**Key Point #2.2: Timestamp Sampling for Continuous Denoisers.** Timestamp sampling of continuous denoisers is also pivotal for  $v_\theta$  to learn the fine-grained knowledge from  $g_\phi$ , *i.e.*,  $I(V; G_1)$ . We compare our proposed *scaled* Logit-Normal sampling with standard uniform sampling, as shown in Table 4. Compared to uniform sampling, ours favors sampling closer to the middle ( $t = 0.5$ ), *i.e.*, in the middle of two distributions  $x_t = tx_1 + (1-t)x_0$ , making denoising more challenging and more beneficial for enhancing  $I(V; G_1)$ . For example, our proposed distribution

Table 4. Comparison of timestamp sampling in continuous denoisers on MMVP-VLM. O: OpenAI CLIP. M: MetaCLIP.

Distribution	Scale	O@224	O@336	M@224
Uniform	N/A	21.5	22.2	23.7
Logit-Normal	0.1	27.4	25.9	26.7
	0.5	28.2	28.9	29.6
	1.0	<b>31.9</b>	<b>29.6</b>	<b>31.9</b>
	5.0	24.5	25.9	25.9
	10.0	20.7	20.0	21.5

Table 5. Performance on SigLIP@224 across different sizes of lightweight continuous and discrete denoisers.

Continuous	#DiT Blocks (MM+Single)	1+2	2+4	3+6	4+8
	MMVP-VLM	41.5	<b>42.2</b>	<b>42.2</b>	41.5
Discrete	#Perceiver Layers	2	4	6	8
	MMVP-VLM	41.5	43.7	<b>45.2</b>	43.7

outperforms uniform sampling by 10.4%, 7.4% and 8.2% on three CLIP backbones in Table 4. Additionally, when the scale  $s$  is too small (e.g.,  $s = 0.1$ , sampling too around 0.5) or too large (e.g.,  $s = 10$ , sampling close to 0 or 1), the lack of diversity in  $t$  can lead to suboptimal results due to the lack of diversity. In this work, we set  $s = 1$  by default.

**Key Point #2.3: Sizes of lightweight denoisers.** We further explore the impact of the size of lightweight denoisers. For the continuous RF, we consider the number of blocks in MM-DiT and Single DiT. We consider the number of layers for Perceiver. Table 5 demonstrates that the denoiser could perform remarkably well with a relatively small size, indicating the efficiency of our lightweight denoisers.

**Key Point #3: Continuous and Discrete Denoisers.** Table 6 demonstrates the performance with continuous and discrete denoisers. Both of them surpass previous work [58] on various backbones. For example, the discrete denoiser obtains a 4.5% performance gain on SigLIP@224 [66]. In summary, our method is general and applies to both continuous and discrete models. It is efficient with lightweight denoisers but strong enough to outperform prior arts [58]. Notably, previous Key Points #1~#2 are consistently applicable to both continuous and discrete denoisers, further highlighting the versatility of our method.

## 5.4. Further Analysis

**Why are improvements on SigLIP relatively small?** In Table 1, we observe that the improvement on SigLIP is relatively smaller compared to OpenAI CLIP and MetaCLIP. Specifically, the performance gain over the original SigLIP is  $\sim 3.7\%$ , less than that for others, i.e.,  $> 10\%$ . Unlike the other two backbones [41, 62], SigLIP [66] does not explicitly train a distinct class token. In practice, we extract the `pooler_output` of SigLIP as the condition for the denoiser, which is obtained by aggregating all local tokens through attention and linear layers. We attribute the relatively small improvement on SigLIP to the indirect leakage

Table 6. Performance of our method with our continuous and discrete denoisers on MMVP-VLM (average of all visual patterns). **Bold** and underline indicate the best and the second best.

Method	OpenAI@224	SigLIP@224	SigLIP@384
DIVA	25.9	40.7	38.5
Continuous	<b>31.9</b>	<u>42.2</u>	<u>40.0</u>
Discrete	<u>28.9</u>	<b>45.2</b>	<b>40.7</b>

Table 7. Efficiency comparison of our lightweight RF denoiser with pre-trained FLUX.1-dev.

Denoiser	Efficiency			MMVP-VLM	
	#Params	Memory	Time/100 iters	OpenAI	Meta-H
Pre-trained	11.90B	37.33G	198.57s	<b>32.6</b>	<b>37.1</b>
Lightweight	<b>1.31B</b>	<b>13.07G</b>	<b>20.55s</b>	31.9	<b>37.1</b>

of local information through the `pooler_output`, which hinders the enhancement of  $I(V; G_1)$ . This is consistent with the discussion in Sec. 4.2 and the results in Fig. 5.

**Efficiency analysis compared with pre-trained FLUX.** We provide a comparison between our lightweight RF ( $n = 2$ ) and original FLUX.1-dev [24] across the following dimensions: #params of denoisers, per-device GPU memory and training time of 100 iterations. To ensure fair comparisons, we fix a per-device batch size of 2. As Table 7 shows, our lightweight denoiser is much more efficient than the pre-trained heavy one. Specifically, our lightweight denoiser has approximately 1/10 of the parameters, occupies about 1/3 of the memory, and is 10 times faster in training, while the final performance remains comparable.

## 6. Conclusive Remarks

In this paper, we delve into the underlying principles of how generative models enhance visual representations. We innovatively uncover that the perfect generation does not always yield optimal representations. The pivot is to learn useful knowledge from the generative model while mitigating irrelevant information. Our key findings lie in three aspects. (1) Conditioning mechanism. We found that local tokens could make the reconstruction task too easy, while class token *alone* as the condition makes the reconstruction task meaningful and significantly enhances visual representations. (2) Denoising configurations. We propose a novel two-stage post-training method to enable vision encoders committed to learning fine-grained knowledge while alleviating irrelevant content. (3) Our model design enables both continuous and discrete denoisers to effectively enhance visual representations. Vision-centric evaluations demonstrate that our method with lightweight denoisers can significantly outperform previous methods relying on heavy pre-trained generative models. We hope this work will inspire further in-depth explorations into the synergy between generative and discriminative models, as well as the relationship between generation and understanding tasks.

## References

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 6
- [3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1:8, 2024. 1
- [4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 7
- [5] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2
- [8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2, 3, 4, 5
- [9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 1, 5
- [10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 3
- [11] Michael Fuest, Pingchuan Ma, Ming Gui, Johannes Schusterbauer, Vincent Tao Hu, and Bjorn Ommer. Diffusion models and representation learning: A survey. *arXiv preprint arXiv:2407.00783*, 2024. 3
- [12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3
- [13] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 3, 6, 7
- [14] Yuxin Guo, Siyang Sun, Shuailei Ma, Kecheng Zheng, Xiaoyi Bao, Shijie Ma, Wei Zou, and Yun Zheng. Cross-mae: Cross-modality masked autoencoders for region-aware audio-visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26721–26731, 2024. 3
- [15] Yuxin Guo, Shuailei Ma, Shijie Ma, Xiaoyi Bao, Chen-Wei Xie, Kecheng Zheng, Tingyu Weng, Siyang Sun, Yun Zheng, and Wei Zou. Aligned better, listen better for audio-visual large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1
- [17] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 5
- [18] Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23115–23127, 2024. 2, 3
- [19] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 4, 5
- [20] Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. In *European Conference on Computer Vision*, pages 113–132. Springer, 2024. 3
- [21] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 4, 5
- [22] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 7
- [23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 7
- [24] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 4, 5, 8
- [25] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 7
- [26] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna,

- Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *Advances in Neural Information Processing Systems*, 37:17044–17068, 2024. 3, 6, 7
- [27] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, 2023. 3, 6, 7
- [28] Haokun Lin, Teng Wang, Yixiao Ge, Yuying Ge, Zhichao Lu, Ying Wei, Qingfu Zhang, Zhenan Sun, and Ying Shan. Toklip: Marry visual tokens to clip for multimodal comprehension and generation. *arXiv preprint arXiv:2505.05422*, 2025. 2
- [29] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 3
- [30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 6
- [31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2, 6, 7
- [32] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 5
- [33] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 3, 6, 7
- [34] Run Luo, Yunshui Li, Longze Chen, Wanwei He, Ting-En Lin, Ziqiang Liu, Lei Zhang, Zikai Song, Hamid Alinejad-Rokny, Xiaobo Xia, Tongliang Liu, Binyuan Hui, and Min Yang. DEEM: Diffusion models serve as the eyes of large language models for image perception. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [35] Shijie Ma, Fei Zhu, Zhun Zhong, Wenzhuo Liu, Xu-Yao Zhang, and Cheng-Lin Liu. Happy: A debiased learning framework for continual generalized category discovery. *Advances in Neural Information Processing Systems*, 37:50850–50875, 2024. 7
- [36] Shijie Ma, Fei Zhu, Zhun Zhong, Xu-Yao Zhang, and Cheng-Lin Liu. Active generalized category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16890–16900, 2024. 3
- [37] Shijie Ma, Fei Zhu, Zhen Cheng, and Xu-Yao Zhang. Towards trustworthy dataset distillation. *Pattern Recognition*, 157:110875, 2025. 3
- [38] Shijie Ma, Fei Zhu, Xu-Yao Zhang, and Cheng-Lin Liu. Protogcd: Unified and unbiased prototype learning for generalized category discovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 7
- [39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 1
- [40] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 5
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6, 8
- [42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3, 5
- [43] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 5
- [44] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 769–778, 2023. 3
- [45] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 3
- [46] Haoru Tan, Sitong Wu, Fei Du, Yukang Chen, Zhibin Wang, Fan Wang, and Xiaojuan Qi. Data pruning via moving-one-sample-out. In *Neural Information Processing Systems (NeurIPS)*, 2023. 7
- [47] Haoru Tan, Sitong Wu, Zhuotao Tian, Yukang Chen, Xiaojuan Qi, and Jiaya Jia. Saco loss: Sample-wise affinity consistency for vision-language pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [48] Haoru Tan, Sitong Wu, Wei Huang, Shizhen Zhao, and Xiaojuan Qi. Data pruning by information maximization. In *International Conference on Learning Representations (ICLR)*, 2025. 7
- [49] Haoru Tan, Sitong Wu, Bo Zhao, Zeke Xie, and XIAOJUAN QI. Diff-in: Data influence estimation with differential approximation, 2025. 3
- [50] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image

- generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2025. 1
- [51] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36:48382–48402, 2023. 3
- [52] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024. 2, 3, 6, 7
- [53] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 2, 3, 6, 7
- [54] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 3
- [55] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3
- [56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3
- [57] Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3
- [58] Wenxuan Wang, Quan Sun, Fan Zhang, Yepeng Tang, Jing Liu, and Xinlong Wang. Diffusion feedback helps CLIP see better. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 5, 6, 7, 8
- [59] Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16284–16294, 2023. 3
- [60] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 7
- [61] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 5
- [62] Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 6, 8
- [63] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023. 1
- [64] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014. 7
- [65] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*, 2025. 1
- [66] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1, 6, 8
- [67] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3