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

Supplementary Material

Overview

In this appendix, we provide additional descriptions of the following contents:

- Relationship with prior works in Appendix A, including some discussions about the differences.
- More training details of hyperparameters in Appendix B.
- Diagrams of various timestamp sampling distributions in Appendix C.
- Additional experimental results in Appendix D.
- Additional qualitative results and cases of the enhanced CLIP (Appendix E) and MLLMs with our enhanced CLIP (Appendix F).
- We also attach algorithms of our two-stage training with continuous and discrete denoisers in Appendix G.

A. Relationship with Prior Works

In this paper, we propose a two-stage post-training method to enhance discriminative models' fine-grained visual representations. For discriminative models, we primarily choose CLIP [8], considering its wide range of applications. Specifically, CLIP is inherently a vision-language model, capable of image-text retrieval and matching. Additionally, CLIP ViT is widely employed as a visual encoder in Multimodal Large Language Models (MLLMs) [4, 5, 11]. Note that our approach follows a post-training paradigm, where we enhance the fine-grained capabilities of a pre-trained CLIP ViT, while preserving its original global semantics.

Comparison with DIVA [14]. DIVA is a pioneering work and proposes to enhance visual representations of CLIP ViT through diffusion feedback. It independently enhances CLIP ViT's visual representations with the guidance of pre-trained stable diffusion [9]. Similar to DIVA, our work focuses on enhancing CLIP ViT's internal visual representations. The enhanced CLIP itself could be a more competent vision-language model with better image-text retrieval performance. Furthermore, the enhanced CLIP ViT serves as a *plug-and-play* module and could be seamlessly plugged into MLLMs. When using the same training recipes but with the enhanced vision encoder, MLLMs could be more capable on several vision-centric benchmarks, with better fine-grained perception on visual details and overcoming visual shortcomings brought about by the original CLIP.

Different from DIVA, we delve into the underlying principles of how generative models enhance vision models [1] from various orthogonal dimensions. Notably, we only employ lightweight denoisers without pre-trained weights of heavy generative models. Our method is efficient yet stronger than DIVA. We also provide several key insights about how to enhance visual representations, *i.e.*, conditioning mechanisms and training configurations. We further explore the implementation of both continuous and discrete generative models. When equipped with corresponding tailor-made designs, both continuous and discrete denoisers outperform DIVA.

Comparison with ROSS [13]. Ross is a pioneering work that explores the intrinsic signals in the vision modality and proposes to append vision-centric self-supervision into the training of MLLMs. The core difference between ROSS and our method is that, ROSS is directly oriented to training better MLLMs. In most cases, ROSS freezes CLIP ViT and enhances the vision-centric performance of MLLMs through the parameters of LLMs. In contrast, our method is directly oriented to enhance CLIP ViT's visual representations. Our method is more general, and the resulting enhanced CLIP could be plugged into various MLLMs. In summary, we independently enhance CLIP ViT, which could be merged into MLLMs for further enhancements, while ROSS directly enhances MLLMs with the ViT frozen.

B. More Training Details

Default training settings. Our training process consists of two stages, each involving one epoch on the CC3M [10] dataset. We choose AdamW [6] 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 [3] with a rank of 16. We train the model on 8 GPUs with a per-device batch size of 16, and the gradient accumulation steps are set as 2, resulting in a global batch size of 256. We plug LoRA to CLIP ViT, with a rank of 16, and an α of 16. Additionally, we employ dropout with a ratio of 0.1 within LoRA.

Detailed settings in Fig. 1 of main manuscript. The default settings are: a lightweight denoiser with 2 MM-DiT [7] and 4 Single-DiT blocks, using only the `[CLS]` as the condition, and two-stage training with 100,000 steps in stage-1 and 5,000 steps in stage-2. Each of the four aspects in Fig. 1(b) modifies *only* one parameter or dimension at a time. Specifically, (i) changes #iters in stage-2 from 100 to 10,000. (ii) varies the number of denoiser-blocks ($n\times$MM-DiT+$2n\times$Single-DiT) from $n = 1$ to $n = 3$. (iii) conditions denoisers with `[CLS]` along with $n\%$ of local tokens. (iv) compares the lightweight denoiser with $n = 4$ and the pretrained heavy `FLUX` with $n = 19$.

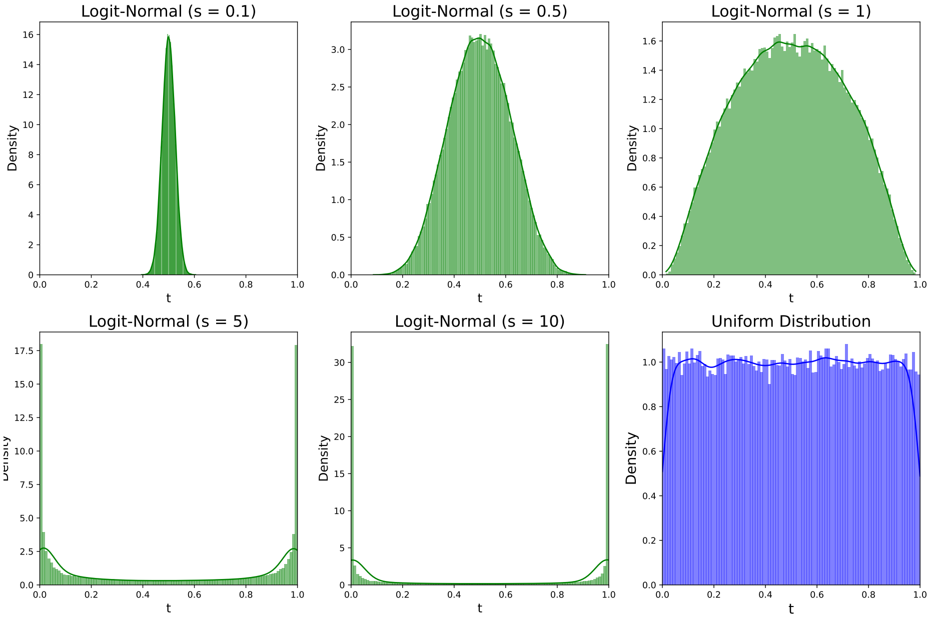


Figure S1. Probability density function of different distributions.

Detailed settings in Table 7 of main manuscript. The pretrained `FLUX` is also trained under the same setting, *i.e.*, two-stage training and *only* the `[CLS]` serves as the condition. Without these proposed keypoints, pretrained denoiser also fails to gain desirable results, *e.g.*, 32.9→22.2 on MMVP, further indicating the generality of our method.

C. Diagrams of Timestamp Sampling

The *scaled* Logit-Normal timestamp sampling [2] is:

$$t = \mathtt{sigmoid}(s \cdot \varepsilon), \quad \text{where } \varepsilon \sim \mathcal{N}(0, 1). \tag{S1}$$

We provide some illustrative diagrams to show the distribution of several candidate distributions, as shown in Fig. S1. In our *scaled* Logit-Normal sampling, as s decreases, the distribution becomes more focused on sampling around the middle ($t = 0.5$). Conversely, as s increases, the distribution becomes more biased towards sampling at the extremes, *i.e.*, $t = 0$ or 1.

D. More Experimental Results

The effect of LoRA. In Stage-2, we apply LoRA to the visual model. The reason is that directly training on the visual model causes rapid updates, which can easily damage the model's high-level semantics and lead to overfitting. By using LoRA, the model can be trained on a larger variety of samples, allowing it to learn more generalizable and fine-grained representations. We conduct experiments on several CLIP backbones, and compare the performance with directly training and LoRA training, as shown in Fig. S2. The performance with LoRA for the visual encoder consistently outperforms the cases of direct training.

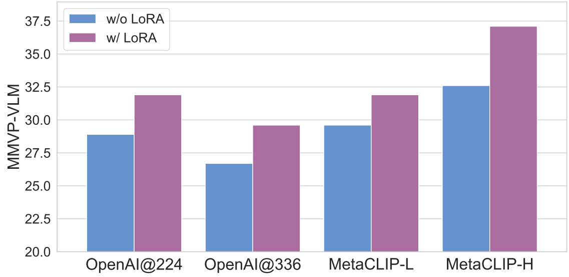


Figure S2. The effect of LoRA on several CLIP backbones.

Table S1. Performance of various mask ratios on OpenAICLIP@224.

Mask Ratio (%)	50	60	70	75	80	85	90	random (50-90)
MMVP-VLM	28.1	27.4	**28.9**	27.4	26.7	25.9	25.9	**28.9**

Whether to update denoiser and projector in Stage-2. In the main text, we argue that in Stage-1, the visual encoder should be fixed, and we train the denoiser and projector. In this way, the projector could learn to bridge the gap between the feature spaces, which serves as the irrelevant information G_2 for visual enhancements. While in Stage-2, we begin to train CLIP ViT to enhance its visual representations. We empirically found that whether the denoiser and projector are updated in Stage-2 has marginal impacts on the final results, as long as Stage-1 training is sufficient. The results are shown in Fig. S4.

Performance with various mask ratios. In the discrete denoiser, we apply masking mechanisms. Here, we provide experimental results across various mask ratios of OpenAICLIP@224, as shown in Table S1.

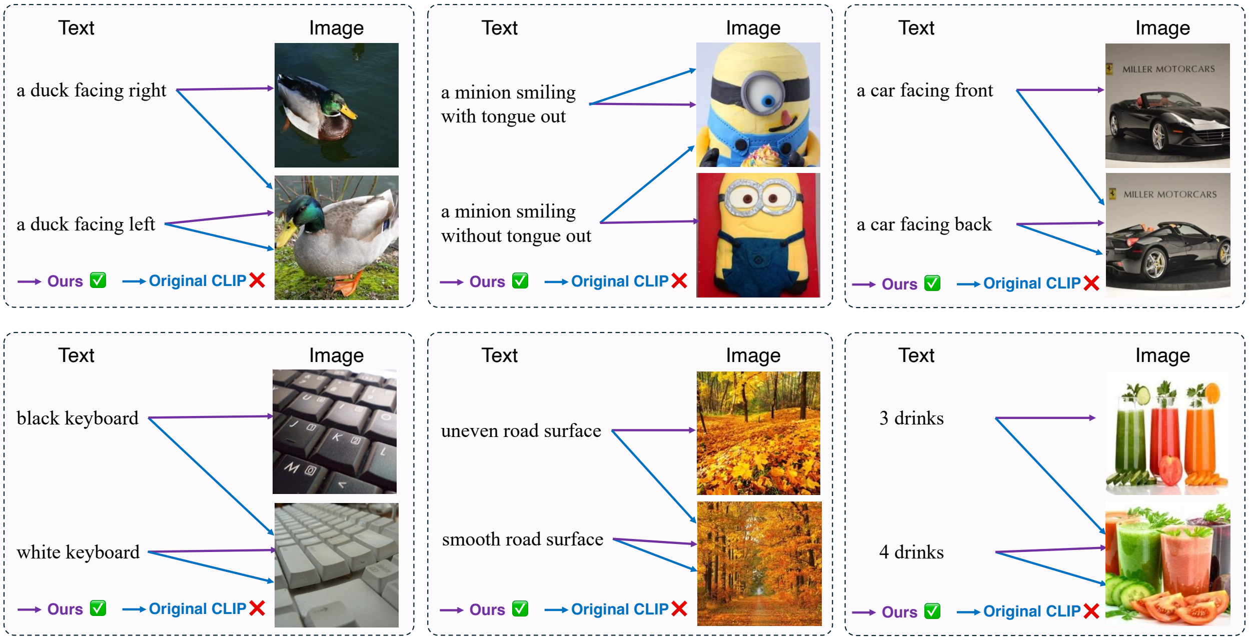


Figure S3. Qualitative results of CLIP. The enhanced CLIP overcomes the original visual shortcomings in fine-grained details.

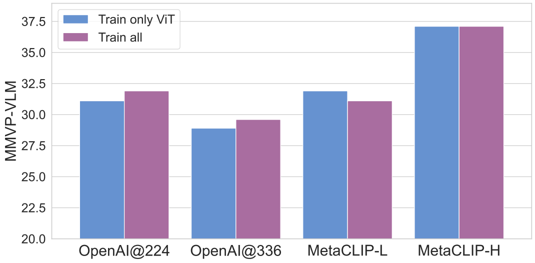


Figure S4. The performance of whether to update the deonoiser and the projector in Stage-2.

E. Qualitative Results of CLIPs

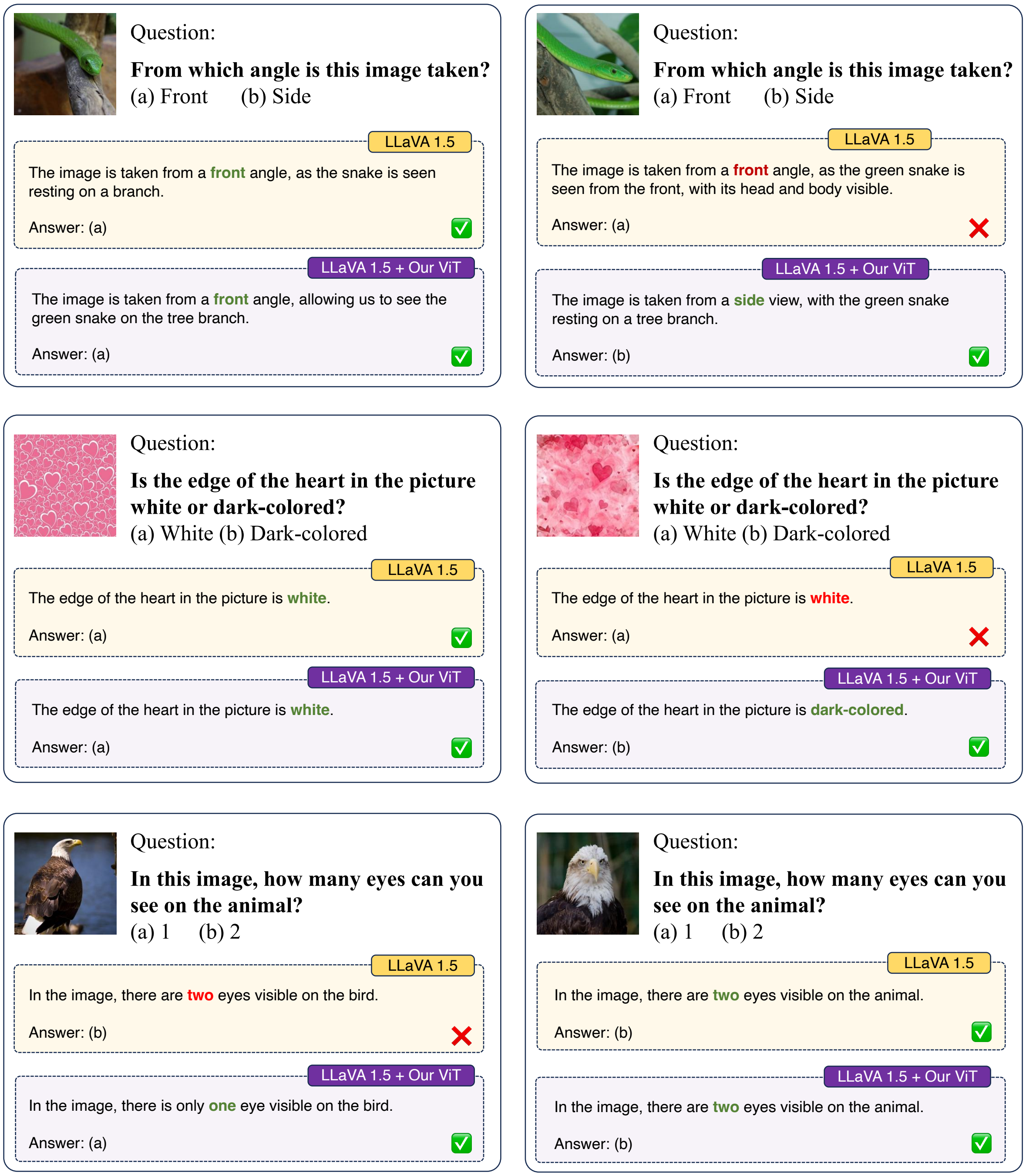
We provide further qualitative results of the original CLIP and our enhanced CLIP, as shown in Fig. S3. The enhanced CLIP overcomes original visual shortcomings in fine-grained details, including color, quantity, structural characteristics, as well as state.

F. Qualitative Results of MLLMs

We provide qualitative results of LLaVA-1.5 [5] with original CLIP ViT and our enhanced CLIP ViT, as shown in Fig S5. Our enhanced visual model could further boost MLLMs' fine-grained visual perception abilities [12].

G. Algorithms

For a clearer and more thorough understanding of our method, we attach the algorithm details of two-stage post-training with continuous and discrete denoisers in Algorithm 1 and Algorithm 2, respectively.



Question:

From which angle is this image taken?
(a) Front (b) Side

LLaVA 1.5

The image is taken from a front angle, as the snake is seen resting on a branch.

Answer: (a) ✓

LLaVA 1.5 + Our ViT

The image is taken from a front angle, allowing us to see the green snake on the tree branch.

Answer: (a) ✓

Question:

From which angle is this image taken?
(a) Front (b) Side

LLaVA 1.5

The image is taken from a front angle, as the green snake is seen from the front, with its head and body visible.

Answer: (a) ✗

LLaVA 1.5 + Our ViT

The image is taken from a side view, with the green snake resting on a tree branch.

Answer: (b) ✓

Question:

Is the edge of the heart in the picture white or dark-colored?
(a) White (b) Dark-colored

LLaVA 1.5

The edge of the heart in the picture is white.

Answer: (a) ✓

LLaVA 1.5 + Our ViT

The edge of the heart in the picture is white.

Answer: (a) ✓

Question:

Is the edge of the heart in the picture white or dark-colored?
(a) White (b) Dark-colored

LLaVA 1.5

The edge of the heart in the picture is white.

Answer: (a) ✗

LLaVA 1.5 + Our ViT

The edge of the heart in the picture is dark-colored.

Answer: (b) ✓

Question:

In this image, how many eyes can you see on the animal?
(a) 1 (b) 2

LLaVA 1.5

In the image, there are two eyes visible on the bird.

Answer: (b) ✗

LLaVA 1.5 + Our ViT

In the image, there is only one eye visible on the bird.

Answer: (a) ✓

Question:

In this image, how many eyes can you see on the animal?
(a) 1 (b) 2

LLaVA 1.5

In the image, there are two eyes visible on the animal.

Answer: (b) ✓

LLaVA 1.5 + Our ViT

In the image, there are two eyes visible on the animal.

Answer: (b) ✓

Figure S5. Qualitative results of MLLMs on MMVP-MLLM benchmark. When equipped with our enhanced CLIP, MLLMs produce better vision-centric performance.

Algorithm 1 Two-stage Visual Enhancements with Continuous Lightweight Denoiser

```
Input: Lightweight and random-initialized denoiser $\boldsymbol{g}_{\phi}(\cdot)$, with lightweight FLUX-like architecture (MM-DiT + Single-DiT).
Input: Pre-trained CLIP ViT $\boldsymbol{v}_{\theta}(\cdot)$ for fine-grained visual representation enhancements.
Input: Random initialized projector $\boldsymbol{h}_{\omega}(\cdot)$ to bridge the feature space of $\boldsymbol{v}_{\theta}$ and condition space of $\boldsymbol{g}_{\phi}$.
Input: The scale hyperparameter $s$ in the proposed *scaled* Logit-Normal sampling.
Input: Pre-trained VAE $\texttt{vae}(\cdot)$ to provide latent space for generative modeling.
Input: Image-only training dataset $\mathcal{D}$ without annotations.
# =================================== Stage-1 ===================================
for $\boldsymbol{x}$ in $\mathcal{D}$ do
    ▷ Prepare input data for generative modeling in latent space:  $\widetilde{\boldsymbol{x}_1} = \texttt{vae}(\boldsymbol{x})$ and $\widetilde{\boldsymbol{x}_0} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.
    ▷ Interpolating in the feature space:  $\widetilde{\boldsymbol{x}_t} = t\widetilde{\boldsymbol{x}_1} + (1-t)\widetilde{\boldsymbol{x}_0}$.
    ▷ Visual encoding as conditions for denoisers:  $\boldsymbol{h}_{\omega} \circ \boldsymbol{v}_{\theta}(\boldsymbol{x})$.
    ▷ Timestamp sampling via *scaled* Logit-Normal distributions:  $\varepsilon \sim \mathcal{N}(0,1)$ then $t = \texttt{sigmoid}(s \cdot \varepsilon)$.
    ▷ Denoising regression objective (flow matching):    # only update $\boldsymbol{g}_{\phi}$ and $\boldsymbol{h}_{\omega}$.
        $$\arg\min_{\phi,\omega} \mathbb{E}_{t,\boldsymbol{x},\widetilde{\boldsymbol{x}}_0,\widetilde{\boldsymbol{x}}_1} \left\| (\widetilde{\boldsymbol{x}_1} - \widetilde{\boldsymbol{x}_0}) - \boldsymbol{g}_{\phi}\left(\widetilde{\boldsymbol{x}_t}, t, \boldsymbol{h}_{\omega} \circ \boldsymbol{v}_{\theta}(\boldsymbol{x})\right) \right\|_2^2.$$
end for
# =================================== Stage-2 ===================================
Plug LoRA upon $\boldsymbol{v}_{\theta}$.
for $\boldsymbol{x}$ in $\mathcal{D}$ do
    ▷ Prepare input data for generative modeling in latent space:  $\widetilde{\boldsymbol{x}_1} = \texttt{vae}(\boldsymbol{x})$ and $\widetilde{\boldsymbol{x}_0} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.
    ▷ Interpolating in the feature space:  $\widetilde{\boldsymbol{x}_t} = t\widetilde{\boldsymbol{x}_1} + (1-t)\widetilde{\boldsymbol{x}_0}$.
    ▷ Visual encoding as conditions for denoisers:  $\boldsymbol{h}_{\omega} \circ \boldsymbol{v}_{\theta}(\boldsymbol{x})$.
    ▷ Timestamp sampling via *scaled* Logit-Normal distributions:  $\varepsilon \sim \mathcal{N}(0,1)$ then $t = \texttt{sigmoid}(s \cdot \varepsilon)$.
    ▷ Denoising regression objective (flow matching):    # update $\boldsymbol{v}_{\theta}$.  Optional:  $\boldsymbol{g}_{\phi}$ and $\boldsymbol{h}_{\omega}$.
        $$\arg\min_{\theta} \mathbb{E}_{t,\boldsymbol{x},\widetilde{\boldsymbol{x}}_0,\widetilde{\boldsymbol{x}}_1} \left\| (\widetilde{\boldsymbol{x}_1} - \widetilde{\boldsymbol{x}_0}) - \boldsymbol{g}_{\phi}\left(\widetilde{\boldsymbol{x}_t}, t, \boldsymbol{h}_{\omega} \circ \boldsymbol{v}_{\theta}(\boldsymbol{x})\right) \right\|_2^2.$$
end for
Output: The enhanced visual model $\boldsymbol{v}_{\theta}^{\star}$ with stronger fine-grained representations.
```

Algorithm 2 Two-stage Visual Enhancements with Discrete Lightweight Denoiser

```
Input: Lightweight and random-initialized denoiser $\boldsymbol{g}_\phi(\cdot)$, instantiated with a lightweight Perceiver.
Input: Pre-trained CLIP ViT $\boldsymbol{v}_\theta(\cdot)$ for fine-grained visual representation enhancements.
Input: Random initialized projector $\boldsymbol{h}_\omega(\cdot)$ to bridge the feature space of $\boldsymbol{v}_\theta$ and condition space of $\boldsymbol{g}_\phi$.
Input: Mask ratio $r$ for discrete modeling.
Input: Pre-trained VQ-GAN $\texttt{vq-gan}(\cdot)$ to discrete indices for generative modeling.
Input: Image-only training dataset $\mathcal{D}$ without annotations.
# ================================== Stage-1 ==================================
for $\boldsymbol{x}$ in $\mathcal{D}$ do
    ▷ Obtain latent embeddings and corresponding discrete indices of input data in VQ-GAN's codebook: $\widetilde{\boldsymbol{x}}, s = \texttt{vq-gan}(\boldsymbol{x})$.
    ▷ Masking $\boldsymbol{x}$'s tokens with ratio $r$ to obtain masked part $\widetilde{\boldsymbol{x}}_{mask}, s_{mask}$ and unmasked part $\widetilde{\boldsymbol{x}}_{unmask}, s_{unmask}$.
    ▷ Visual encoding and obtain conditions via cross-attention for denoisers:
        $Q = \widetilde{\boldsymbol{x}}_{unmask},$
        $K, V = \texttt{concat}\big(\widetilde{\boldsymbol{x}}_{unmask}; \boldsymbol{h}_\omega \circ \boldsymbol{v}_\theta(\boldsymbol{x})\big),$
        $\boldsymbol{c}_{\omega,\theta} = \texttt{cross-attn}(Q, K, V).$
    ▷ Denoising cross-entropy objective (masked index prediction):    # only update $\boldsymbol{g}_\phi$ and $\boldsymbol{h}_\omega$.
        $\arg\min_{\phi,\omega} \mathbb{E}_{\boldsymbol{x}} - \log \prod_{i=1}^{L} \boldsymbol{g}_\phi\big(s_{mask} | s_{unmask}, \boldsymbol{c}_{\omega,\theta}\big).$
end for
# ================================== Stage-2 ==================================
Plug LoRA upon $\boldsymbol{v}_\theta$.
for $\boldsymbol{x}$ in $\mathcal{D}$ do
    ▷ Obtain latent embeddings and corresponding discrete indices of input data in VQ-GAN's codebook: $\widetilde{\boldsymbol{x}}, s = \texttt{vq-gan}(\boldsymbol{x})$.
    ▷ Masking $\boldsymbol{x}$'s tokens with ratio $r$ to obtain masked part $\widetilde{\boldsymbol{x}}_{mask}, s_{mask}$ and unmasked part $\widetilde{\boldsymbol{x}}_{unmask}, s_{unmask}$.
    ▷ Visual encoding and obtain conditions via cross-attention for denoisers:
        $Q = \widetilde{\boldsymbol{x}}_{unmask},$
        $K, V = \texttt{concat}\big(\widetilde{\boldsymbol{x}}_{unmask}; \boldsymbol{h}_\omega \circ \boldsymbol{v}_\theta(\boldsymbol{x})\big),$
        $\boldsymbol{c}_{\omega,\theta} = \texttt{cross-attn}(Q, K, V).$
    ▷ Denoising cross-entropy objective (masked index prediction):    # update $\boldsymbol{v}_\theta$. Optional: $\boldsymbol{g}_\phi$ and $\boldsymbol{h}_\omega$.
        $\arg\min_{\theta} \mathbb{E}_{\boldsymbol{x}} - \log \prod_{i=1}^{L} \boldsymbol{g}_\phi\big(s_{mask} | s_{unmask}, \boldsymbol{c}_{\omega,\theta}\big).$
end for
Output: The enhanced visual model $\boldsymbol{v}_\theta^\star$ with stronger fine-grained representations.
```

References

[1] 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. 1

[2] 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. 2

[3] 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. 1

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[5] 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. 1, 3

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1

[7] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 1

[8] 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

[9] 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

[10] 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. 1

[11] 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. 1

[12] 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. 3

[13] 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. 1

[14] 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. 1