Guiding Diffusion-based Reconstruction with Contrastive Signals for Balanced Visual Representation

Supplementary Material

Contents

A. Symbol Definitions

In this section, Table 1 includes a summary of key notations and descriptions in this work.

Table 1. A summary of key notations and descriptions in this work.

Notations	Descriptions
$\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$	Input image.
$\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$	Training dataset.
$\tilde{\mathbf{x}}$	VAE latent of image $\mathbf{x}$.
$t \in \{1, \dots, T\}$	Diffusion timestep.
$\mathbf{x}_t$	Noisy latent at diffusion step t.
ϵ_θ	Diffusion noise prediction network.
f_ϕ	CLIP visual encoder.
$\mathbf{z}$	Visual feature extracted by CLIP encoder, shorthand for $\mathbf{z} = f_\phi(\mathbf{x})$.
h_ω	Projection module that maps CLIP features to diffusion condition space.
c	Condition used for diffusion denoising, shorthand for $c = h_\omega(\mathbf{z})$.
$\hat{\boldsymbol{\epsilon}}_\theta$	Predicted noise, shorthand for $\hat{\boldsymbol{\epsilon}}_\theta = \epsilon_\theta(\tilde{\mathbf{x}}_t, c, t)$.
$\boldsymbol{\epsilon}_t^{\text{gt}}$	Ground-truth diffusion noise at step t.
$\hat{\boldsymbol{\epsilon}}_+/\hat{\boldsymbol{\epsilon}}_-$	Predicted noise of positive/negative sample.
P	Positive set, shorthand for $P = \{\hat{\boldsymbol{\epsilon}}_+, \boldsymbol{\epsilon}_{\text{gt}}\}$.
N	Negative set, shorthand for $N = \{\hat{\boldsymbol{\epsilon}}_-^k\}_{k=1}^{N-1}$.
C	Set $C = P \cup N$.
$\text{sim}(u, v)$	Cosine similarity $\text{sim}(u, v) = \frac{\langle u,v\rangle}{\lvert u\rvert\cdot\lvert v\rvert}$.

B. Algorithm of DCR

For clarity, we provide the pseudocode of our *Diffusion Contrastive Reconstruction (DCR)* training procedure in Algorithm 1.

Algorithm 1: DCR Algorithm.

Input: Dataset $\mathcal{D} = \{\mathbf{x}_i\}$; CLIP vision encoder f_ϕ; projector h_ω; frozen diffusion model ϵ_θ; VAE encoder Enc; diffusion steps T; batch size N_b; temperature τ; Stage-1 epochs E_1; Stage-2 epochs E_2;

Output: Enhanced CLIP encoder f_ϕ;

▷ Stage-1: Projector Alignment
for $epoch = 1$ **to** E_1 **do**
  Sample mini-batch $\{\mathbf{x}_i\}_{i=1}^{N_b}$ from $\mathcal{D}$;
  for $i = 1$ **to** N_b **do**
    Sample timestep $t \sim \text{Unif}(1, T)$ and noise $\boldsymbol{\epsilon}_t^{\text{gt}} \sim \mathcal{N}(0, I)$;
    $\tilde{\mathbf{x}}_i \leftarrow \text{Enc}(\mathbf{x}_i)$;
    $\mathbf{x}_t \leftarrow \sqrt{\bar{\alpha}_t}\,\tilde{\mathbf{x}}_i + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_t^{\text{gt}}$;
    $\hat{\boldsymbol{\epsilon}}_i \leftarrow \epsilon_\theta(\mathbf{x}_t, h_\omega(f_\phi(\mathbf{x}_i)), t)$;
    $\mathbf{x}_i^+ \leftarrow a(\mathbf{x}_i)$;
    $\hat{\boldsymbol{\epsilon}}_{+,i} \leftarrow \epsilon_\theta(\mathbf{x}_t, h_\omega(f_\phi(\mathbf{x}_i^+)), t)$;
    $N_i \leftarrow \emptyset$;
    for $j = 1$ **to** $N_b, j \neq i$ **do**
      $\hat{\boldsymbol{\epsilon}}_{-,i}^j \leftarrow \epsilon_\theta(\mathbf{x}_t, h_\omega(f_\phi(\mathbf{x}_j)), t)$;
      $N_i \leftarrow N_i \cup \{\hat{\boldsymbol{\epsilon}}_{-,i}^j\}$;
    end for
    $P_i \leftarrow \{\hat{\boldsymbol{\epsilon}}_{+,i}, \boldsymbol{\epsilon}_t^{\text{gt}}\}$;
    $C_i \leftarrow P_i \cup N_i$;
    $\mathcal{L}_i \leftarrow -\frac{1}{2}\sum_{p\in P_i} \log \frac{\exp(\text{sim}(\hat{\boldsymbol{\epsilon}}_i, p)/\tau)}{\sum_{c\in C_i}\exp(\text{sim}(\hat{\boldsymbol{\epsilon}}_i, c)/\tau)}$;
  end for
  $\mathcal{L}_{dcr} \leftarrow \frac{1}{N_b}\sum_{i=1}^{N_b}\mathcal{L}_i$;
  Update projector parameters ω;
end for
▷ Stage-2: Encoder Enhancement
for $epoch = 1$ **to** E_2 **do**
  Sample mini-batch $\{\mathbf{x}_i\}_{i=1}^{N_b}$ and repeat the same steps to compute P_i, N_i, and $\mathcal{L}_i$;
  $\mathcal{L}_{dcr} \leftarrow \frac{1}{N_b}\sum_{i=1}^{N_b}\mathcal{L}_i$;
  Update CLIP vision encoder f_ϕ;
end for
return f_ϕ;

C. Proof of Theorem 1

Definition 1 (Bi-Lipschitz Mapping [3])**.** Let $T : \mathcal{X} \to \mathcal{Y}$ be a mapping between two metric spaces $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ and $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$. We say that T is bi-Lipschitz on a subset $\mathcal{M} \subseteq \mathcal{X}$ if there exist constants $0 < m \leq L < \infty$ such that, for all $z_1, z_2 \in \mathcal{M}$,

$$m\|z_1 - z_2\|_{\mathcal{X}} \leq \big\|T(z_1) - T(z_2)\big\|_{\mathcal{Y}} \leq L\|z_1 - z_2\|_{\mathcal{X}}. \quad (1)$$

In particular, a bi-Lipschitz map preserves pairwise distances up to constant factors and is injective on $\mathcal{M}$.

Assumption 1. Fix a diffusion step t and define $T(\mathbf{z}) = \epsilon_\theta\big(x_t, h_\omega(\mathbf{z}), t\big)$. According to Definition 1, we assume that $T(\mathbf{z})$ satisfies the bi-Lipschitz property.

Lemma 1 (Equivalent Variance Identity [22])**.** *Let $e_1, \ldots, e_n \in \mathbb{R}^d$ be arbitrary vectors and denote their mean by $\bar{e} = \frac{1}{n}\sum_{i=1}^n e_i$. Then the empirical variance can be written as the average of all pairwise squared distances:*

$$\sum_{i=1}^n \|e_i - \bar{e}\|_2^2 = \frac{1}{2n}\sum_{i=1}^n\sum_{j=1}^n \|e_i - e_j\|_2^2. \quad (2)$$

Proof. **For the left-hand side**, using $\bar{e} = \frac{1}{n}\sum_{k=1}^n e_k$, we have:

$$\begin{aligned}\sum_{i=1}^n \|e_i - \bar{e}\|_2^2 &= \sum_{i=1}^n \langle e_i - \bar{e}, e_i - \bar{e}\rangle \\ &= \sum_{i=1}^n \big(\|e_i\|_2^2 + \|\bar{e}\|_2^2 - 2\langle e_i, \bar{e}\rangle\big) \\ &= \sum_{i=1}^n \|e_i\|_2^2 + n\|\bar{e}\|_2^2 - 2\left\langle \sum_{i=1}^n e_i, \bar{e}\right\rangle.\end{aligned} \quad (3)$$

Since $\sum_{i=1}^n e_i = n\bar{e}$, it follows that

$$\left\langle \sum_{i=1}^n e_i, \bar{e}\right\rangle = \langle n\bar{e}, \bar{e}\rangle = n\|\bar{e}\|_2^2. \quad (4)$$

Therefore,

$$\sum_{i=1}^n \|e_i - \bar{e}\|_2^2 = \sum_{i=1}^n \|e_i\|_2^2 - n\|\bar{e}\|_2^2. \quad (5)$$

For the right-hand side, using $\|e_i - e_j\|_2^2 = \|e_i\|_2^2 + \|e_j\|_2^2 - 2\langle e_i, e_j\rangle$, we obtain

$$\begin{aligned}&\sum_{i=1}^n\sum_{j=1}^n \|e_i - e_j\|_2^2 \\ =&\sum_{i,j} \|e_i\|_2^2 + \sum_{i,j}\|e_j\|_2^2 - 2\sum_{i,j}\langle e_i, e_j\rangle \\ =&n\sum_{i=1}^n \|e_i\|_2^2 + n\sum_{j=1}^n \|e_j\|_2^2 - 2\left\langle \sum_{i=1}^n e_i, \sum_{j=1}^n e_j\right\rangle \\ =&2n\sum_{i=1}^n \|e_i\|_2^2 - 2\left\|\sum_{i=1}^n e_i\right\|_2^2.\end{aligned} \quad (6)$$

Again using $\sum_{i=1}^n e_i = n\bar{e}$, we obtain

$$\left\|\sum_{i=1}^n e_i\right\|_2^2 = \|n\bar{e}\|_2^2 = n^2\|\bar{e}\|_2^2, \quad (7)$$

so

$$\sum_{i=1}^n\sum_{j=1}^n \|e_i - e_j\|_2^2 = 2n\sum_{i=1}^n \|e_i\|_2^2 - 2n^2\|\bar{e}\|_2^2. \quad (8)$$

Dividing Eq. (8) by $2n$ yields

$$\frac{1}{2n}\sum_{i=1}^n\sum_{j=1}^n \|e_i - e_j\|_2^2 = \sum_{i=1}^n \|e_i\|_2^2 - n\|\bar{e}\|_2^2. \quad (9)$$

Therefore, by combining Eq. (5) and Eq. (9), we can obtain:

$$\sum_{i=1}^n \|e_i - \bar{e}\|_2^2 = \frac{1}{2n}\sum_{i=1}^n\sum_{j=1}^n \|e_i - e_j\|_2^2. \quad (10)$$

This completed the proof. □

Restate of Theorem 1. Fix a diffusion step t and define $T(\mathbf{z}) = \epsilon_\theta\big(x_t, h_\omega(\mathbf{z}), t\big)$. Under a mild assumption, the intra-class scatter and inter-class scatter in the feature space $(S_{\text{inner}}, S_{\text{inter}})$ can be bounded by those in the noise space $(S_{\text{inner}}^{(\epsilon)}(t), S_{\text{inter}}^{(\epsilon)}(t))$ as

$$S_{\text{inner}} \leq \frac{1}{m^2} S_{\text{inner}}^{(\epsilon)}(t),$$

$$S_{\text{inter}} \geq \kappa S_{\text{inter}}^{(\epsilon)}(t) - \eta S_{\text{inner}}^{(\epsilon)}(t),$$

where m, κ, and η are positive constants depending on the Lipschitz continuity of $T(\cdot)$.

Proof. We proceed in three steps. Throughout, fix a diffusion step t and a mini-batch with class set $\mathcal{Y}$.

For class $y \in \mathcal{Y}$, let the feature set be $\mathcal{Z}_y = \{\mathbf{z}_i\}_{i=1}^{n_y}$ with $\mathbf{z}_i = f_\phi(\mathbf{x}_i)$ and the noise set be $\mathcal{E}_{y,t} = \{\hat{\boldsymbol{\epsilon}}_i\}_{i=1}^{n_y}$ with

$\hat{\boldsymbol{\epsilon}}_i = \epsilon_\theta(x_t, h_\omega(\mathbf{z}_i), t)$. Denote the class means by $\mu_y = \frac{1}{n_y}\sum_{i=1}^{n_y}\mathbf{z}_i$ and $\nu_{y,t} = \frac{1}{n_y}\sum_{i=1}^{n_y}\hat{\boldsymbol{\epsilon}}_i$. As shown in Eq. (4) and Eq. (5) in the main text, we define the intra-/inter-class scatters in the CLIP feature space by

$$S_{\text{inner}} = \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\frac{1}{n_y}\sum_{i=1}^{n_y}\|\mathbf{z}_i - \mu_y\|_2^2, \tag{11}$$

$$S_{\text{inter}} = \frac{1}{|\mathcal{Y}|(|\mathcal{Y}|-1)}\sum_{\substack{y,y'\in\mathcal{Y}\\ y\neq y'}}\|\mu_y - \mu_{y'}\|_2^2. \tag{12}$$

Similarly, we define the intra-/inter-class scatters in the diffusion noise space by

$$S^{(\epsilon)}_{\text{inner}}(t) = \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\frac{1}{n_y}\sum_{i=1}^{n_y}\|\hat{\boldsymbol{\epsilon}}_i - \nu_{y,t}\|_2^2, \tag{13}$$

$$S^{(\epsilon)}_{\text{inter}}(t) = \frac{1}{|\mathcal{Y}|(|\mathcal{Y}|-1)}\sum_{\substack{y,y'\in\mathcal{Y}\\ y\neq y'}}\|\nu_{y,t} - \nu_{y',t}\|_2^2. \tag{14}$$

Step 1: Minimizing $\mathcal{L}_{\text{DCR}}$ improves noise-space geometry.

Recall the DCR loss

$$\mathcal{L}_{\text{DCR}} = -\frac{1}{2}\sum_{p\in P}\log\frac{d(\hat{\boldsymbol{\epsilon}}, p)}{\sum_{c\in C} d(\hat{\boldsymbol{\epsilon}}, c)},$$

where $\hat{\boldsymbol{\epsilon}} = \epsilon_\theta(x_t,\, h_\omega(\mathbf{z}),\, t)$ is the anchor, $\hat{\boldsymbol{\epsilon}}_+$ is the positive sample, $\boldsymbol{\epsilon}_{\text{gt}}$ is the ground-truth, and $N = \{\hat{\boldsymbol{\epsilon}}_-^{(j)}\}$ are negatives from other images in the mini-batch. $P = \{\hat{\boldsymbol{\epsilon}}_+,\, \boldsymbol{\epsilon}_t^{\text{gt}}\}$, $C = P \cup N$. $d(u, v) = \exp(\text{sim}(u, v)/\tau)$ with sim the cosine similarity and $\tau > 0$. Let

$$p(q \mid \hat{\boldsymbol{\epsilon}}) = \frac{\exp(\text{sim}(\hat{\boldsymbol{\epsilon}}, q)/\tau)}{\sum_{c\in C}\exp(\text{sim}(\hat{\boldsymbol{\epsilon}}, c)/\tau)}. \tag{15}$$

Differentiating the loss with respect to the similarity term gives the following gradient expression:

$$\frac{\partial\mathcal{L}_{\text{DCR}}}{\partial\,\text{sim}(\hat{\boldsymbol{\epsilon}}, q)} = \begin{cases} -\frac{1}{2\tau}\big(1 - 2p(q \mid \hat{\boldsymbol{\epsilon}})\big), & q \in P, \\ \frac{1}{\tau}\,p(q \mid \hat{\boldsymbol{\epsilon}}), & q \in N. \end{cases} \tag{16}$$

Thus, gradient descent on $\mathcal{L}_{\text{DCR}}$ **increases** the anchor's cosine similarity with the two positives ($\hat{\boldsymbol{\epsilon}}_+$ and $\boldsymbol{\epsilon}_{\text{gt}}$) and **decreases** it with all negatives. Since

$$\text{sim}(u, v) = \langle\bar{u}, \bar{v}\rangle, \quad \text{where } \bar{u} = \frac{u}{\|u\|},\ \bar{v} = \frac{v}{\|v\|}, \tag{17}$$

we have

$$\|\bar{u} - \bar{v}\|_2^2 = \|\bar{u}\|_2^2 + \|\bar{v}\|_2^2 - 2\langle\bar{u}, \bar{v}\rangle = 2(1 - \text{sim}(\bar{u}, \bar{v})). \tag{18}$$

Therefore, minimizing $\mathcal{L}_{\text{DCR}}$ is equivalent to

$$\min_{\phi,\omega}\big\|\hat{\boldsymbol{\epsilon}} - \hat{\boldsymbol{\epsilon}}_+\big\|_2^2 + \big\|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\text{gt}}\big\|_2^2, \tag{19}$$

$$\max_{\phi,\omega}\big\|\hat{\boldsymbol{\epsilon}} - \hat{\boldsymbol{\epsilon}}_-^{(j)}\big\|_2^2,\ \forall j. \tag{20}$$

Then, we pass from pairwise distances to scatter. Specifically, for each class $y \in \mathcal{Y}$, the key component in Eq. (13),

$$\frac{1}{n_y}\sum_{i=1}^{n_y}\|\hat{\boldsymbol{\epsilon}}_i - \nu_{y,t}\|_2^2 \tag{21}$$

is, by Lemma 1, exactly the average of all pairwise squared distances $\|\hat{\boldsymbol{\epsilon}}_i - \hat{\boldsymbol{\epsilon}}_j\|_2^2$ within that class (up to the constant factor $1/(2n_y^2)$). Minimizing the pairwise distances between the anchor and its positive/ground-truth terms in Eq. (19) therefore decreases the average intra-class pairwise distance, and thus decreases $\frac{1}{n_y}\sum_{i=1}^{n_y}\|\hat{\boldsymbol{\epsilon}}_i - \nu_{y,t}\|_2^2$ for each y. Averaging over all classes yields

$$\min_{\phi,\omega}\ S^{(\epsilon)}_{\text{inner}}(t). \tag{22}$$

In contrast, the negative terms $\{\hat{\boldsymbol{\epsilon}}_-^{(j)}\}$ are drawn from other images in the mini-batch, which typically belong to different classes. Maximizing the anchor–negative distances $\|\hat{\boldsymbol{\epsilon}} - \hat{\boldsymbol{\epsilon}}_-^{(j)}\|_2^2$ increases the average pairwise distances between samples of different classes. This increase, when propagated to the distances between the corresponding class means $\{\nu_{y,t}\}_{y\in\mathcal{Y}}$, leads to an increase of the inter-class scatter $S^{(\epsilon)}_{\text{inter}}(t)$:

$$\max_{\phi,\omega}\ S^{(\epsilon)}_{\text{inter}}(t). \tag{23}$$

Combining Eq. (22) and Eq. (23), we conclude that minimizing $\mathcal{L}_{\text{DCR}}$ improves the geometry in the noise space by decreasing intra-class scatter and increasing inter-class scatter.

Step 2: Bounding feature-space intra-class scatter.

Fix a class $y \in \mathcal{Y}$, and write its feature set as $\mathcal{Z}_y = \{\mathbf{z}_i\}_{i=1}^{n_y}$ and its noise set as $\mathcal{E}_{y,t} = \{\hat{\boldsymbol{\epsilon}}_i\}_{i=1}^{n_y}$ with $\hat{\boldsymbol{\epsilon}}_i = T(\mathbf{z}_i)$.

By Lemma 1, the intra-class variance in the feature space can be written as

$$\frac{1}{n_y}\sum_{i=1}^{n_y}\|\mathbf{z}_i - \mu_y\|_2^2 = \frac{1}{2n_y^2}\sum_{i=1}^{n_y}\sum_{j=1}^{n_y}\|\mathbf{z}_i - \mathbf{z}_j\|_2^2. \tag{24}$$

Similarly, the intra-class variance in the noise space is

$$\frac{1}{n_y}\sum_{i=1}^{n_y}\|\hat{\boldsymbol{\epsilon}}_i - \nu_{y,t}\|_2^2 = \frac{1}{2n_y^2}\sum_{i=1}^{n_y}\sum_{j=1}^{n_y}\|\hat{\boldsymbol{\epsilon}}_i - \hat{\boldsymbol{\epsilon}}_j\|_2^2. \tag{25}$$

By Assumption 1, there exists $m > 0$ such that, for all $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{M}$,

$$m\|\mathbf{z}_1 - \mathbf{z}_2\| \le \|T(\mathbf{z}_1) - T(\mathbf{z}_2)\|. \tag{26}$$

Equivalently,

$$\|\mathbf{z}_1 - \mathbf{z}_2\| \le \frac{1}{m}\|T(\mathbf{z}_1) - T(\mathbf{z}_2)\|. \tag{27}$$

Applying this inequality to each pair $(\mathbf{z}_i, \mathbf{z}_j)$ gives

$$\|\mathbf{z}_i - \mathbf{z}_j\|^2 \le \frac{1}{m^2}\|T(\mathbf{z}_i) - T(\mathbf{z}_j)\|^2 = \frac{1}{m^2}\|\hat{\boldsymbol{\epsilon}}_i - \hat{\boldsymbol{\epsilon}}_j\|^2. \tag{28}$$

Substituting Eq. (28) into Eq. (24) and comparing with Eq. (25), we obtain

$$\begin{aligned}\frac{1}{n_y}\sum_{i=1}^{n_y}\|\mathbf{z}_i - \mu_y\|_2^2 &\le \frac{1}{2n_y^2}\sum_{i,j=1}^{n_y}\frac{1}{m^2}\|\hat{\boldsymbol{\epsilon}}_i - \hat{\boldsymbol{\epsilon}}_j\|_2^2 \\ &= \frac{1}{m^2}\cdot\frac{1}{n_y}\sum_{i=1}^{n_y}\|\hat{\boldsymbol{\epsilon}}_i - \nu_{y,t}\|_2^2.\end{aligned} \tag{29}$$

Averaging over all classes $y \in \mathcal{Y}$ yields

$$\begin{aligned}S_{\text{inner}} &= \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\frac{1}{n_y}\sum_{i=1}^{n_y}\|\mathbf{z}_i - \mu_y\|_2^2 \\ &\le \frac{1}{m^2}\cdot\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\frac{1}{n_y}\sum_{i=1}^{n_y}\|\hat{\boldsymbol{\epsilon}}_i - \nu_{y,t}\|_2^2 \\ &= \frac{1}{m^2}S_{\text{inner}}^{(\epsilon)}(t),\end{aligned} \tag{30}$$

which proves the first inequality in Theorem 1.

Step 3: Bounding feature-space inter-class scatter.

By Assumption 1, there exists $L > 0$ such that, for all $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{M}$,

$$\|T(\mathbf{z}_1) - T(\mathbf{z}_2)\| \le L\|\mathbf{z}_1 - \mathbf{z}_2\|. \tag{31}$$

Equivalently,

$$\|\mathbf{z}_1 - \mathbf{z}_2\| \ge \frac{1}{L}\|T(\mathbf{z}_1) - T(\mathbf{z}_2)\|. \tag{32}$$

Fix two distinct classes $y, y' \in \mathcal{Y}$. Applying Eq. (32) with $\mathbf{z}_1 = \mu_y$, $\mathbf{z}_2 = \mu_{y'}$ yields

$$\|\mu_y - \mu_{y'}\| \ge \frac{1}{L}\|T(\mu_y) - T(\mu_{y'})\|. \tag{33}$$

We next relate $T(\mu_y)$ to $\nu_{y,t}$. We have

$$\nu_{y,t} = \frac{1}{n_y}\sum_{i=1}^{n_y}\hat{\boldsymbol{\epsilon}}_i = \frac{1}{n_y}\sum_{i=1}^{n_y}T(\mathbf{z}_i). \tag{34}$$

Then

$$\begin{aligned}\|T(\mu_y) - \nu_{y,t}\|_2 &= \left\|T(\mu_y) - \frac{1}{n_y}\sum_{i=1}^{n_y}T(\mathbf{z}_i)\right\|_2 \\ &= \left\|\frac{1}{n_y}\sum_{i=1}^{n_y}\big(T(\mu_y) - T(\mathbf{z}_i)\big)\right\|_2 \\ &\le \frac{1}{n_y}\sum_{i=1}^{n_y}\|T(\mu_y) - T(\mathbf{z}_i)\|_2 \\ &\le \frac{L}{n_y}\sum_{i=1}^{n_y}\|\mu_y - \mathbf{z}_i\|_2,\end{aligned} \tag{35}$$

where we used the triangle inequality and then the Lipschitz property of T. By the Cauchy-Schwarz inequality,

$$\sum_{i=1}^{n_y}\|\mu_y - \mathbf{z}_i\|_2 \le \sqrt{n_y}\left(\sum_{i=1}^{n_y}\|\mu_y - \mathbf{z}_i\|_2^2\right)^{1/2}. \tag{36}$$

Hence,

$$\|T(\mu_y) - \nu_{y,t}\|_2 \le \frac{L}{\sqrt{n_y}}\left(\sum_{i=1}^{n_y}\|\mu_y - \mathbf{z}_i\|_2^2\right)^{1/2}. \tag{37}$$

Similarly, for class y', we have:

$$\|T(\mu_{y'}) - \nu_{y',t}\|_2 \le \frac{L}{\sqrt{n_{y'}}}\left(\sum_{i=1}^{n_{y'}}\|\mu_{y'} - \mathbf{z}_i\|_2^2\right)^{1/2}. \tag{38}$$

Using the triangle inequality, we get

$$\begin{aligned}&\|T(\mu_y) - T(\mu_{y'})\|_2 \\ \ge&\|\nu_{y,t} - \nu_{y',t}\|_2 - \|T(\mu_y) - \nu_{y,t}\|_2 \\ &- \|T(\mu_{y'}) - \nu_{y',t}\|_2.\end{aligned} \tag{39}$$

Combining this with Eqs. (37) and (38), we obtain

$$\begin{aligned}&\|T(\mu_y) - T(\mu_{y'})\|_2 \\ \ge&\|\nu_{y,t} - \nu_{y',t}\|_2 - \frac{L}{\sqrt{n_y}}\left(\sum_{i=1}^{n_y}\|\mu_y - \mathbf{z}_i\|_2^2\right)^{1/2} \\ &- \frac{L}{\sqrt{n_{y'}}}\left(\sum_{i=1}^{n_{y'}}\|\mu_{y'} - \mathbf{z}_i\|_2^2\right)^{1/2}.\end{aligned} \tag{40}$$

Substituting Eq. (40) into Eq. (33), we obtain

$$\begin{aligned}\|\mu_y - \mu_{y'}\|_2 \ge \frac{1}{L}&\|\nu_{y,t} - \nu_{y',t}\|_2 \\ &- \frac{1}{\sqrt{n_y}}\left(\sum_{i=1}^{n_y}\|\mu_y - \mathbf{z}_i\|_2^2\right)^{1/2} \\ &- \frac{1}{\sqrt{n_{y'}}}\left(\sum_{i=1}^{n_{y'}}\|\mu_{y'} - \mathbf{z}_i\|_2^2\right)^{1/2}.\end{aligned} \tag{41}$$

Squaring both sides and using the inequality $(a-b-c)^2 \geq \frac{1}{2}a^2 - 2(b^2+c^2)$ for any $a, b, c \in \mathbb{R}$, we obtain

$$\begin{aligned}\|\mu_y - \mu_{y'}\|_2^2 \geq \frac{1}{2L^2}\|\nu_{y,t} - \nu_{y',t}\|_2^2 \\ - 2\left(\frac{1}{n_y}\sum_{i=1}^{n_y}\|\mu_y - \mathbf{z}_i\|_2^2\right) \\ - 2\left(\frac{1}{n_{y'}}\sum_{i=1}^{n_{y'}}\|\mu_{y'} - \mathbf{z}_i\|_2^2\right).\end{aligned} \tag{42}$$

Averaging the above inequality over all ordered pairs (y, y') with $y \neq y'$, and recalling the definition of S_{inter} and $S_{\text{inter}}^{(\epsilon)}(t)$, we obtain

$$\begin{aligned}S_{\text{inter}} &= \frac{1}{|\mathcal{Y}|(|\mathcal{Y}|-1)}\sum_{\substack{y,y'\in\mathcal{Y}\\ y\neq y'}}\|\mu_y - \mu_{y'}\|_2^2 \\ &\geq \frac{1}{|\mathcal{Y}|(|\mathcal{Y}|-1)}\sum_{\substack{y,y'\in\mathcal{Y}\\ y\neq y'}}\left[\frac{1}{2L^2}\|\nu_{y,t} - \nu_{y',t}\|_2^2 - 2B_{y,y'}\right] \\ &= \frac{1}{2L^2}S_{\text{inter}}^{(\epsilon)}(t) - \frac{2}{|\mathcal{Y}|(|\mathcal{Y}|-1)}\sum_{\substack{y,y'\in\mathcal{Y}\\ y\neq y'}}B_{y,y'},\end{aligned} \tag{43}$$

where $B_{y,y'} = \frac{1}{n_y}\sum_{i\in I_y}\|\mu_y - \mathbf{z}_i\|_2^2 + \frac{1}{n_{y'}}\sum_{i\in I_{y'}}\|\mu_{y'} - \mathbf{z}_i\|_2^2$.

For the second term, note that for each fixed y, the quantity $A_y = \frac{1}{n_y}\sum_{i\in I_y}\|\mu_y - \mathbf{z}_i\|_2^2$ appears exactly $(|\mathcal{Y}|-1)$ times as the first index and $(|\mathcal{Y}|-1)$ times as the second index when summing over all ordered pairs (y, y') with $y \neq y'$. Hence

$$\sum_{\substack{y,y'\in\mathcal{Y}\\ y\neq y'}}(A_y + A_{y'}) = 2(|\mathcal{Y}|-1)\sum_{y\in\mathcal{Y}}A_y. \tag{44}$$

Therefore,

$$\begin{aligned}&\frac{2}{|\mathcal{Y}|(|\mathcal{Y}|-1)}\sum_{\substack{y,y'\in\mathcal{Y}\\ y\neq y'}}(A_y + A_{y'}) \\ =&\frac{2}{|\mathcal{Y}|(|\mathcal{Y}|-1)}\,2(|\mathcal{Y}|-1)\sum_{y\in\mathcal{Y}}A_y \\ =&\frac{4}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}A_y \\ =&4S_{\text{inner}}.\end{aligned} \tag{45}$$

Substituting this back, we obtain the explicit bound

$$S_{\text{inter}} \geq \frac{1}{2L^2}S_{\text{inter}}^{(\epsilon)}(t) - 4S_{\text{inner}}. \tag{46}$$

Finally, using the bound from Step 2, $S_{\text{inner}} \leq \frac{1}{m^2}S_{\text{inner}}^{(\epsilon)}(t)$, we arrive at

$$S_{\text{inter}} \geq \frac{1}{2L^2}S_{\text{inter}}^{(\epsilon)}(t) - \frac{4}{m^2}S_{\text{inner}}^{(\epsilon)}(t). \tag{47}$$

Thus, we have:

$$S_{\text{inter}} \geq \kappa S_{\text{inter}}^{(\epsilon)}(t) - \eta S_{\text{inner}}^{(\epsilon)}(t), \tag{48}$$

where $\kappa = \frac{1}{2L^2}$ and $\eta = \frac{4}{m^2}$. It proves the second inequality in Theorem 1.

This completed the proof. □

D. Proof of Theorem 2

Restate of Theorem 2. If negatives in the mini-batch are well separated from the anchor (*i.e.*, $\text{sim}(\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\epsilon}}_-^j) \ll \text{sim}(\hat{\boldsymbol{\epsilon}}, \boldsymbol{\epsilon}_{\text{gt}})$ for all $j \neq i$), and predicted-noise norms are bounded away from 0 and ∞, then the DCR loss reduces to a scaled reconstruction loss up to an additive constant:

$$\mathcal{L}_{\text{dcr}} = \lambda\left\|\epsilon_\theta(\mathbf{x}_t,\ h_\omega(f_\phi(\tilde{\mathbf{x}})), t) - \boldsymbol{\epsilon}_t^{\text{gt}}\right\|_2^2 + c,$$

where $\lambda > 0$, and c is a constant.

Proof. We consider a single anchor sample and drop the sample index for brevity.

Recall that the anchor is $\hat{\boldsymbol{\epsilon}} = \epsilon_\theta(x_t,\ h_\omega(f_\phi(\tilde{\mathbf{x}})),\ t)$, the two positives are $\hat{\boldsymbol{\epsilon}}_+$ and $\boldsymbol{\epsilon}_{\text{gt}}$, and $N = \{\hat{\boldsymbol{\epsilon}}_-^j\}_j$ denotes the set of negatives in the mini-batch. Let $P = \{\hat{\boldsymbol{\epsilon}}_+, \boldsymbol{\epsilon}_{\text{gt}}\}$, $C = P \cup N$, and $\text{sim}(u, v) = \frac{\langle u,v\rangle}{\|u\|_2\,\|v\|_2}$, $d(u, v) = \exp\big(\text{sim}(u, v)/\tau\big)$, with $\tau > 0$ the temperature. The DCR loss for this anchor is

$$\mathcal{L}_{\text{DCR}} = -\frac{1}{2}\sum_{p\in P}\log\frac{d(\hat{\boldsymbol{\epsilon}}, p)}{\sum_{c\in C}d(\hat{\boldsymbol{\epsilon}}, c)}. \tag{49}$$

Denote $u = \text{sim}(\hat{\boldsymbol{\epsilon}}, \boldsymbol{\epsilon}_{\text{gt}})$, $v = \text{sim}(\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\epsilon}}_+)$, and write

$$Z_{\text{pos}} = \sum_{p\in P}\exp\big(\text{sim}(\hat{\boldsymbol{\epsilon}}, p)/\tau\big) = e^{u/\tau} + e^{v/\tau}, \tag{50}$$

$$Z_{\text{neg}} = \sum_j \exp\big(\text{sim}(\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\epsilon}}_-^j)/\tau\big). \tag{51}$$

Then we can expand

$$\begin{aligned}\mathcal{L}_{\text{DCR}} &= -\frac{1}{2}\sum_{p\in P}\left(\frac{\text{sim}(\hat{\boldsymbol{\epsilon}}, p)}{\tau} - \log\big(Z_{\text{pos}} + Z_{\text{neg}}\big)\right) \\ &= -\frac{1}{2\tau}(u+v) + \log\big(Z_{\text{pos}} + Z_{\text{neg}}\big).\end{aligned} \tag{52}$$

Using the notation in Eq. (52), we now show that $\mathcal{L}_{\text{DCR}}$ is (up to scaling and an additive constant) equivalent to the reconstruction loss.

First, by the assumption that negatives are well separated from the anchor, there exist constants $\Delta > 0$ and $B \in \mathbb{N}$ such that for all anchors in the mini-batch,

$$\mathrm{sim}(\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\epsilon}}_-^j) \le u - \Delta, \quad \forall j, \qquad |N| \le B, \tag{53}$$

where $u = \mathrm{sim}(\hat{\boldsymbol{\epsilon}}, \boldsymbol{\epsilon}_{\text{gt}})$. Hence

$$Z_{\text{neg}} = \sum_j \exp\big(\mathrm{sim}(\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\epsilon}}_-^j)/\tau\big) \le B \exp\big((u-\Delta)/\tau\big). \tag{54}$$

Since $Z_{\text{pos}} \ge e^{u/\tau}$, we get

$$\frac{Z_{\text{neg}}}{Z_{\text{pos}}} \le B e^{-\Delta/\tau} = \delta, \tag{55}$$

where $\delta > 0$ is a constant independent of the trainable parameters. Therefore

$$\begin{aligned} &\log\big(Z_{\text{pos}} + Z_{\text{neg}}\big) \\ &= \log\Big(Z_{\text{pos}}\big(1 + Z_{\text{neg}}/Z_{\text{pos}}\big)\Big) \\ &= \log Z_{\text{pos}} + \log\big(1 + Z_{\text{neg}}/Z_{\text{pos}}\big), \end{aligned} \tag{56}$$

and, using $0 \le Z_{\text{neg}}/Z_{\text{pos}} \le \delta$,

$$0 \le \log\big(1 + Z_{\text{neg}}/Z_{\text{pos}}\big) \le \log(1+\delta) = C_{\text{neg}}. \tag{57}$$

Thus, from Eq. (52),

$$\mathcal{L}_{\text{DCR}} = -\frac{1}{2\tau}(u+v) + \log Z_{\text{pos}} + \Delta_{\text{neg}}, \tag{58}$$

where $0 \le \Delta_{\text{neg}} \le C_{\text{neg}}$. It remains to analyze the positive part:

$$\ell_{\text{pos}}(u,v) = -\frac{1}{2\tau}(u+v) + \log\big(e^{u/\tau} + e^{v/\tau}\big), \tag{59}$$

since

$$\mathcal{L}_{\text{DCR}} = \ell_{\text{pos}}(u,v) + \Delta_{\text{neg}}. \tag{60}$$

Next, we express u and v in terms of squared Euclidean distances. Define the normalized vectors

$$\hat{e} = \frac{\hat{\boldsymbol{\epsilon}}}{\|\hat{\boldsymbol{\epsilon}}\|_2}, \qquad g = \frac{\boldsymbol{\epsilon}_{\text{gt}}}{\|\boldsymbol{\epsilon}_{\text{gt}}\|_2}, \qquad p = \frac{\hat{\boldsymbol{\epsilon}}_+}{\|\hat{\boldsymbol{\epsilon}}_+\|_2}. \tag{61}$$

Then $u = \langle \hat{e}, g\rangle$ and $v = \langle \hat{e}, p\rangle$. For unit vectors a, b we have

$$\|a - b\|_2^2 = 2(1 - \langle a, b\rangle), \tag{62}$$

so, defining

$$d_{\text{gt}}^2 = \|\hat{e} - g\|_2^2, \qquad d_+^2 = \|\hat{e} - p\|_2^2, \tag{63}$$

we obtain

$$u = 1 - \frac{1}{2}d_{\text{gt}}^2, \qquad v = 1 - \frac{1}{2}d_+^2. \tag{64}$$

Substituting into ℓ_{pos} gives

$$\ell_{\text{pos}} = \frac{1}{4\tau}\big(d_{\text{gt}}^2 + d_+^2\big) + \log\Big(e^{-d_{\text{gt}}^2/(2\tau)} + e^{-d_+^2/(2\tau)}\Big). \tag{65}$$

Since $\hat{e}, g, p$ are unit vectors, both squared distances are bounded:

$$0 \le d_{\text{gt}}^2 \le 4, \qquad 0 \le d_+^2 \le 4. \tag{66}$$

The logarithmic term in Eq. (65) satisfies a bound. Let

$$a = -\frac{d_{\text{gt}}^2}{2\tau}, \qquad b = -\frac{d_+^2}{2\tau}, \tag{67}$$

so that

$$\log\Big(e^{-d_{\text{gt}}^2/(2\tau)} + e^{-d_+^2/(2\tau)}\Big) = \log(e^a + e^b). \tag{68}$$

For any real a, b, we have

$$\max\{a,b\} \le \log(e^a + e^b) \le \max\{a,b\} + \log 2. \tag{69}$$

Using $0 \le d_{\text{gt}}^2, d_+^2 \le 4$ and $\tau > 0$, we obtain constants $C_{\min}, C_{\max} \in \mathbb{R}$ depending only on τ such that

$$C_{\min} \le \log\Big(e^{-d_{\text{gt}}^2/(2\tau)} + e^{-d_+^2/(2\tau)}\Big) \le C_{\max} \tag{70}$$

for all possible d_{gt}^2, d_+^2. Moreover, the term $\frac{1}{4\tau}d_+^2$ in Eq. (65) is also bounded using Eq. (66), so it can be absorbed into the constants. Consequently, there exist finite constants $\tilde{C}_{\min}, \tilde{C}_{\max}$ (depending only on τ) such that

$$\frac{1}{4\tau}d_{\text{gt}}^2 + \tilde{C}_{\min} \le \ell_{\text{pos}} \le \frac{1}{4\tau}d_{\text{gt}}^2 + \tilde{C}_{\max}. \tag{71}$$

We now relate d_{gt}^2 to the true reconstruction error $\|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\text{gt}}\|_2^2$. For any nonzero vectors u, v, the law of cosines gives

$$\|u - v\|_2^2 = \|u\|_2^2 + \|v\|_2^2 - 2\|u\|_2\|v\|_2 \,\mathrm{sim}(u,v). \tag{72}$$

With $u = \hat{\boldsymbol{\epsilon}}$, $v = \boldsymbol{\epsilon}_{\text{gt}}$ and $\mathrm{sim}(u,v) = \langle \hat{e}, g\rangle = 1 - d_{\text{gt}}^2/2$, we obtain

$$\begin{aligned} \big\|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\text{gt}}\big\|_2^2 &= \|\hat{\boldsymbol{\epsilon}}\|_2^2 + \|\boldsymbol{\epsilon}_{\text{gt}}\|_2^2 - 2\|\hat{\boldsymbol{\epsilon}}\|_2\|\boldsymbol{\epsilon}_{\text{gt}}\|_2\left(1 - \frac{d_{\text{gt}}^2}{2}\right) \\ &= \big(\|\hat{\boldsymbol{\epsilon}}\|_2 - \|\boldsymbol{\epsilon}_{\text{gt}}\|_2\big)^2 + \|\hat{\boldsymbol{\epsilon}}\|_2\|\boldsymbol{\epsilon}_{\text{gt}}\|_2\, d_{\text{gt}}^2. \end{aligned} \tag{73}$$

By assumption, predicted-noise norms are bounded away from 0 and ∞. Thus, there exist constants $0 < \alpha \le \beta < \infty$ such that

$$\alpha \le \|\hat{\boldsymbol{\epsilon}}\|_2,\ \|\boldsymbol{\epsilon}_{\text{gt}}\|_2 \le \beta \quad \text{for all anchors.} \tag{74}$$

Using Eq. (73) and the fact that $\big(\|\hat{\boldsymbol{\epsilon}}\|_2 - \|\boldsymbol{\epsilon}_{\text{gt}}\|_2\big)^2 \ge 0$, we have

$$\big\|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\text{gt}}\big\|_2^2 \ge \|\hat{\boldsymbol{\epsilon}}\|_2\|\boldsymbol{\epsilon}_{\text{gt}}\|_2\, d_{\text{gt}}^2 \ge \alpha^2 d_{\text{gt}}^2. \tag{75}$$

On the other hand, using $\|\hat{\boldsymbol{\epsilon}}\|_2^2 + \|\boldsymbol{\epsilon}_{\text{gt}}\|_2^2 \le 2\beta^2$ and $\|\hat{\boldsymbol{\epsilon}}\|_2\|\boldsymbol{\epsilon}_{\text{gt}}\|_2 \le \beta^2$, we obtain

$$\left\|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\text{gt}}\right\|_2^2 \le 2\beta^2 + \beta^2 d_{\text{gt}}^2. \tag{76}$$

Rearranging these inequalities yields

$$\frac{1}{\beta^2}\left(\left\|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\text{gt}}\right\|_2^2 - 2\beta^2\right) \le d_{\text{gt}}^2 \le \frac{1}{\alpha^2}\left\|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\text{gt}}\right\|_2^2. \tag{77}$$

Combining Eq. (71) and Eq. (77), we get affine bounds on ℓ_{pos} in terms of the reconstruction error. For the lower bound,

$$\begin{aligned}\ell_{\text{pos}} &\ge \frac{1}{4\tau} d_{\text{gt}}^2 + \tilde{C}_{\min} \\ &\ge \frac{1}{4\tau\beta^2}\left(\left\|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\text{gt}}\right\|_2^2 - 2\beta^2\right) + \tilde{C}_{\min} \\ &= \frac{1}{4\tau\beta^2}\left\|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\text{gt}}\right\|_2^2 + \left(\tilde{C}_{\min} - \frac{1}{2\tau}\right).\end{aligned} \tag{78}$$

For the upper bound,

$$\begin{aligned}\ell_{\text{pos}} &\le \frac{1}{4\tau} d_{\text{gt}}^2 + \tilde{C}_{\max} \\ &\le \frac{1}{4\tau\alpha^2}\left\|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\text{gt}}\right\|_2^2 + \tilde{C}_{\max}.\end{aligned} \tag{79}$$

Define

$$\lambda_{\min} = \frac{1}{4\tau\beta^2}, \qquad \lambda_{\max} = \frac{1}{4\tau\alpha^2}, \tag{80}$$

and constants

$$c_{\min} = \tilde{C}_{\min} - \frac{1}{2\tau}, \qquad c_{\max} = \tilde{C}_{\max}. \tag{81}$$

Then

$$\lambda_{\min}\left\|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\text{gt}}\right\|_2^2 + c_{\min} \le \ell_{\text{pos}} \le \lambda_{\max}\left\|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\text{gt}}\right\|_2^2 + c_{\max}. \tag{82}$$

Finally, recalling Eq. (60) and the bound $0 \le \Delta_{\text{neg}} \le C_{\text{neg}}$, we conclude that

$$\lambda_{\min}\left\|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\text{gt}}\right\|_2^2 + c_{\min} \le \mathcal{L}_{\text{DCR}} \le \lambda_{\max}\left\|\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}_{\text{gt}}\right\|_2^2 + c_{\max} + C_{\text{neg}}. \tag{83}$$

Thus, $\mathcal{L}_{\text{DCR}}$ is sandwiched between two affine functions of the reconstruction loss $\left\|\epsilon_\theta(x_t, h_\omega(f_\phi(\tilde{\mathbf{x}})), t) - \boldsymbol{\epsilon}_t^{\text{gt}}\right\|_2^2$ with strictly positive slopes. In other words, minimizing $\mathcal{L}_{\text{DCR}}$ is equivalent (up to a positive scaling factor and an additive constant) to minimizing the standard reconstruction loss. We can summarize this equivalence in the form

$$\mathcal{L}_{\text{DCR}} = \lambda\left\|\epsilon_\theta(x_t, h_\omega(f_\phi(\tilde{\mathbf{x}})), t) - \boldsymbol{\epsilon}_t^{\text{gt}}\right\|_2^2 + c, \tag{84}$$

for some $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ and constant c.

This completed the proof. □

E. Additional Experimental Settings

In this section, we make a supplementation to Sec 5.1.

E.1. CLIP Backbones

Here, we provide a detailed summary of the 6 types of CLIP backbones used in our experiments. They fall into 3 categories:

- **OpenAI CLIP ViT-L@224 and OpenAI CLIP ViT-L@336** [16] establish the canonical formulation of contrastive language-image pretraining, defining the feature space that later CLIP variants build upon. Their large-scale but noisy training data gives them strong open-world generalization, while their ability to perceive fine-grained visual details remains limited. The 336-resolution variant uniquely extends CLIP's receptive field without altering the architecture, making it a standard choice for studying high-resolution alignment.
- **MetaCLIP ViT-L@224 and MetaCLIP ViT-H@224** [20] distinguishes itself by reconstructing a high-quality, well-aligned training corpus using a principled filtering pipeline rather than relying on raw web data. This dataset-centric redesign leads to representations that are more stable, less noisy, and better calibrated than those of OpenAI CLIP. The ViT-H model further pushes scaling laws within the CLIP paradigm.
- **SigLIP ViT-SO@224 and SigLIP ViT-SO@384** [21] departs from the classic softmax contrastive loss by introducing a sigmoid-based objective, fundamentally altering how positive and negative pairs influence training. This probabilistic formulation enables more fine-grained pairwise alignment and reduces overconfidence artifacts commonly observed in CLIP-like models. Its higher-resolution variant leverages the smoother objective to maintain stable training.

E.2. Competitors

Here we give a more detailed summary of the competitors mentioned in the experiments.

- **Original CLIP** [16, 20, 21] is the baseline vision encoder pretrained on large-scale image–text pairs. It provides generalization across diverse recognition and retrieval tasks. However, it lacks mechanisms for reconstructive feedback, which limits its ability to understand fine-grained visuals.
- **DIVA** [19] introduces diffusion-based visual feedback to refine CLIP features. It performs reconstruction conditioned on CLIP vision embeddings, enabling the model to recover more detailed visual information.
- **GenHancer** [15] systematically investigates how generative models can enhance CLIP by refining conditioning design, denoising strategies, and generation paradigms. It shows that using global conditions and lightweight de-

noisers leads to more stable reconstruction-based representation learning. In addition, it extends the reconstruction process to discrete latent spaces.

- **un^{2}CLIP** [13] builds on the unCLIP framework by inverting the generative process so that the visual encoder can better capture fine-grained image details while remaining aligned with CLIP's original embedding space.

E.3. Evaluation Protocol for P-Ability

Datasets. Following [13, 15, 19], we use MMVP-VLM [18] to evaluate the **P-Ability**. It is a benchmark designed to evaluate fine-grained visual perception by testing VLMs on a wide range of visual patterns. It contains human-designed symbols, shapes, transformations, and composites that isolate specific perceptual abilities such as symmetry, color matching, and geometric relations. The dataset includes tens of thousands of diverse pattern–label pairs, providing a controlled environment for assessing detailed visual understanding.

Evaluation metrics. Following [18], we report Accuracy (ACC) as the evaluation metric. ACC measures the proportion of samples for which the model correctly recognizes the underlying visual pattern, serving as a direct indicator of fine-grained visual perception.

E.4. Evaluation Protocol for D-Ability

Datasets. Following [16], we perform zero-shot clustering on 6 standard datasets [2, 4, 5, 7–9] to evaluate the **D-Ability**.

- **MNIST** [9] is a handwritten digit recognition dataset containing grayscale images of digits 0-9. Each image is centered and normalized to 28×28 pixels, making it a standard benchmark for evaluating basic visual representations. It provides 60000 training and 10000 test samples.
- **CIFAR-10** [8] is a natural image dataset designed for general object classification across 10 categories. The images are low-resolution 32×32 color images that introduce significant appearance variation despite their small size. The dataset includes 50000 training and 10000 test images.
- **Eurosat** [7] is a satellite image dataset for land-use and land-cover classification, containing 10 classes such as agricultural areas, forests, and urban regions. Derived from Sentinel-2 satellite imagery, it includes 21600 training and 5400 test images, offering rich spatial and spectral diversity.
- **Caltech-101** [5] designed for general object classification, includes 101 object categories and a background class, featuring around 7650 training and 3300 test images. The images, collected at Caltech, present significant variation in scale, orientation, and lighting conditions.
- **Describable Textures Dataset (DTD)** [2] focuses on texture classification, featuring 47 texture categories described using human-interpretable attributes. It offers 3760 training and 1880 test images, providing a unique challenge in recognizing visually distinctive patterns from natural and artificial sources.
- **ImageNet-1K** [4] is a large-scale benchmark for visual recognition, covering 1000 object categories spanning animals, scenes, and man-made objects. The images exhibit substantial diversity in viewpoint, background, resolution, and object appearance, making it a core testbed for evaluating high-capacity visual models. It contains about 1.28 million training images and 50 thousand validation images.

Evaluation metrics. Following [12], we adopt 3 widely-used metrics to evaluate the clustering performance.

- **Normalized Mutual Information (NMI)** measures the mutual dependence between predicted clusters and ground-truth labels, normalized to $[0, 1]$. Higher values indicate better alignment between the clustering structure and the true class partition.
- **Accuracy (ACC)** evaluates the best-matched assignment between predicted clusters and ground-truth classes. It computes the fraction of correctly assigned samples after optimal label permutation using the Hungarian algorithm.
- **Adjusted Rand Index (ARI)** quantifies the similarity between two partitions by counting pairwise agreements, adjusted for random chance. An ARI of 0 corresponds to random clustering, while higher values indicate more consistent partitioning.

E.5. Evaluation Protocol for MLLMs

Following [13, 15, 18], we evaluate MLLMs using two complementary groups of benchmarks: **Vision-Centric Benchmarks**, which focus on fine-grained visual perception and scene understanding, and **Conventional MLLM Benchmarks**, which measure robustness, hallucination resistance, and multimodal reasoning. Together, they provide a comprehensive view of an MLLM's visual competence, reliability, and general multimodal capability.

Vision-Centric Benchmarks. We evaluate on 4 datasets:

- **MMVP-MLLM** [18] evaluates the model's fine-grained visual perception by testing its ability to identify structured visual patterns. We report classification accuracy (ACC).
- **NaturalBench** [10] examines multimodal reasoning across natural images using four metrics: overall accuracy (Acc), question-type accuracy (Q-Acc), instance-level accuracy (I-Acc), and group-level accuracy (G-Acc).
- **CV-Bench 2D** [17] measures perception quality on 2D dense prediction tasks. We report the accuracy (ACC) of ADE20K [23] and COCO [1].
- **CV-Bench 3D** [17] assesses 3D spatial understanding

through structured 3D reasoning queries. We report overall accuracy (ACC).

Conventional MLLM Benchmarks. We evaluate on 3 datasets:

- **POPE** [11] evaluates object hallucination robustness under three settings, including random (rand), popularity-based (pop), and adversarial (adv). Higher scores indicate stronger resistance to hallucination.
- **SciQA-IMG** [14] assesses scientific visual question answering involving diagrams, plots, and structured visual cues. We report accuracy (ACC).
- **Hallusion** [6] measures multimodal robustness by testing whether the model avoids visually induced false inferences. We report the average accuracy across all query types.

F. Additional Experimental Results

F.1. Expanded Version of Quantitative Results

Here, we present an expanded version of the quantitative results.

Tab. 2 evaluates the improved CLIP models on the two classical tasks, zero-shot image classification and zero-shot text and image retrieval. The competitor focuses only on fine-grained reconstruction, which neglects discriminative ability and leads to performance drops. In contrast, our method preserves and further improves discriminative ability, demonstrating its effectiveness.

Table 2. Performance on zero-shot classification and retrieval.

Method	Classification						Retrieval-Image@5		Retrieval-Text@5	
	MNIST	C10	Eur	C101	DTD	IN-1K	Flickr30k	COCO	Flickr30k	COCO
Original	76.4	**95.6**	60.1	86.6	**55.4**	75.5	87.2	**61.1**	**97.4**	79.2
Genhancer	69.7	73.7	58.5	71.5	48.4	73.8	81.6	51.1	87.3	61.4
Ours	**76.5**	**95.6**	**60.3**	**86.7**	**55.4**	**75.6**	**87.3**	**61.1**	97.3	**79.3**

Tab. 3 reports the results under different ratios of local tokens (`[CLS]` + $n\%$ local tokens). We find that using too many local tokens leads to performance degradation. This may be because excessive local tokens provide overly strong local cues, making the reconstruction task too easy and thus weakening the supervision signal. This phenomenon is consistent with the findings in GenHancer [15] and further supports the common behavior of reconstruction-based enhancement methods.

Table 3. Results under different ratios of local tokens (`[CLS]` + $n\%$ local tokens).

Ratio	MMVP-VLM	Clustering		
	ACC	NMI	ACC	ARI
0%	**33.30**	**0.76**	**0.67**	**0.54**
10%	31.85	**0.76**	0.63	0.53
50%	23.70	0.73	0.60	0.46
80%	21.48	0.69	0.57	0.44
100%	20.74	0.65	0.55	0.42

F.2. Expanded Version of Qualitative Results

Here, we present an expanded version of the qualitative results.

Fig. 1 presents some qualitative examples from the MMVP-VLM benchmark. The results show that our method enhances fine-grained visual perception, leading to improved **P-Ability**.

Fig. 2 shows several qualitative examples on the vision-centric benchmarks from MLLMs. The results demonstrate that our enhanced CLIP can be seamlessly integrated into MLLMs to improve their visual capabilities.

F.3. Computational Costs

Tab. 4 reports the computational costs of our method. Across different CLIP backbones, the training cost of DCR remains stable and mainly scales with the backbone size and input resolution. Models with similar parameter counts, such as OpenAI CLIP ViT-L@224, OpenAI CLIP ViT-L@336, and MetaCLIP ViT-L@224, exhibit almost identical training times, indicating that DCR introduces minimal overhead beyond the backbone's forward passes. Larger or higher-resolution models, including MetaCLIP ViT-H@224 and SigLIP ViT-SO@384, incur predictable increases in training time and GPU memory usage, yet the overall computation remains well within a practical range for NVIDIA-A100 80GB GPU. These results show that DCR is computationally lightweight and can be seamlessly applied to a wide range of CLIP architectures without requiring specialized optimization.

Table 4. Training costs of our DCR.

CLIP Backbone	Training #Params	Training Time	Training GPU Memory
OpenAI CLIP ViT-L@224	350.5M	7.7h	53544M
OpenAI CLIP ViT-L@336	350.5M	7.6h	60472M
MetaCLIP ViT-L@224	350.5M	7.7h	60472M
MetaCLIP ViT-H@224	360.0M	9.2h	76430M
SigLIP ViT-SO@224	358.5M	8.0h	66798M
SigLIP ViT-SO@384	358.5M	9.7h	78792M

G. Future Works

In the future, we plan to extend our diffusion contrastive reconstruction beyond diffusion models to VAR-based generators, enabling a broader family of generative priors to contribute fine-grained visual supervision. In parallel, we aim to investigate how our framework can be applied to strengthen visual encoders from diverse architectures, thereby providing a unified and adaptable enhancement strategy for a wide range of vision systems. Moreover, we believe that a systematic study of reconstruction-based methods is essential for uncovering the underlying principles of representation enhancement and for establishing rigorous theoretical guarantees.

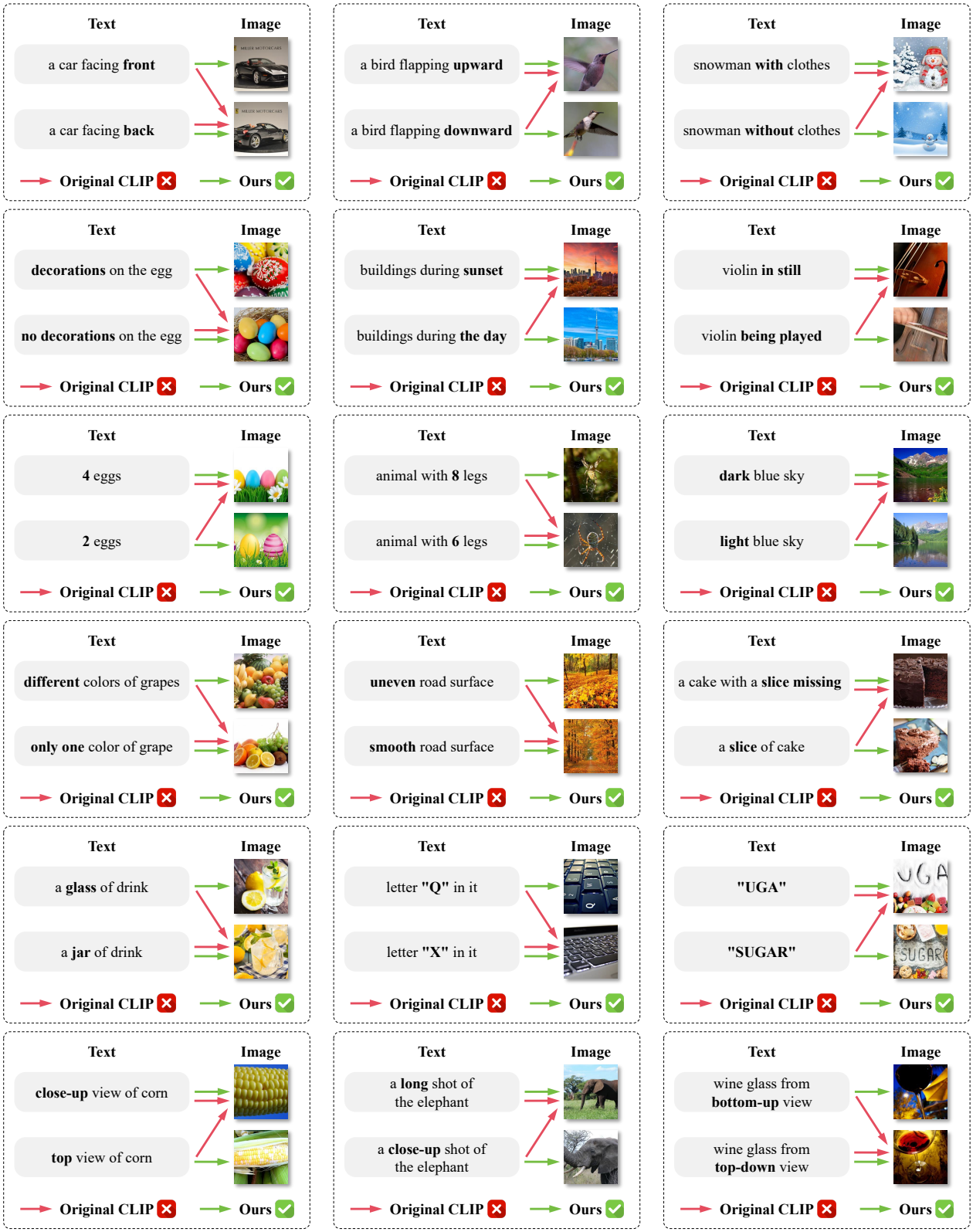
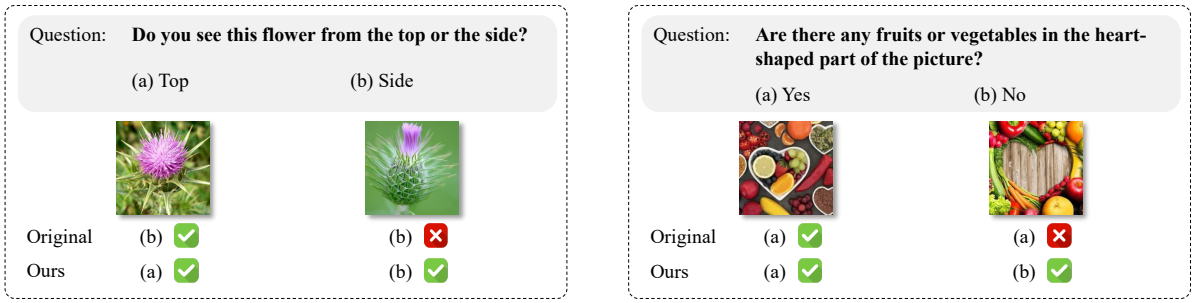
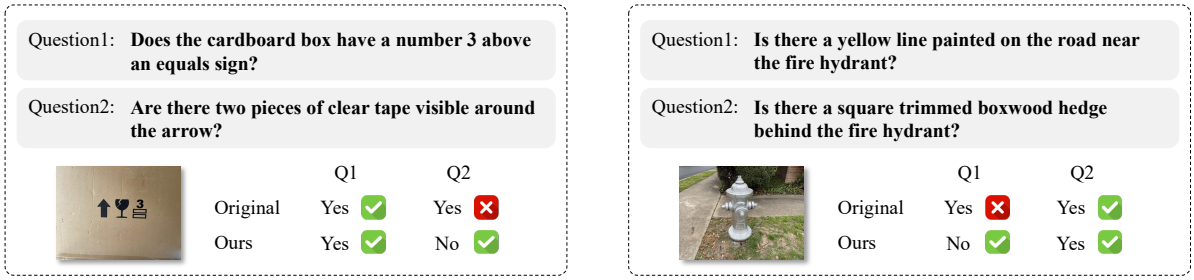


Figure 1. Expanded version of qualitative results on the MMVP-VLM benchmark. The predictions from the original CLIP and our improved version are indicated by red and green arrows, respectively. The improved CLIP effectively addresses the original model's limitations in capturing fine-grained visual details.



(a) MMVP-MLLM.



(b) NaturalBench.

Figure 2. Expanded version of qualitative results on the MLLM benchmarks.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018. 8

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 8

[3] Guy C David. Bi-lipschitz pieces between manifolds. *Revista Mathematica Iberoamericana*, 32(1), 2016. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 8

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR workshop*, pages 178–178, 2004. 8

[6] 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 *CVPR*, pages 14375–14385, 2024. 9

[7] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *JSTARS*, 12(7):2217–2226, 2019. 8

[8] A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009. 8

[9] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002. 8

[10] 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. In *NeurIPS*, pages 17044–17068, 2024. 8

[11] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pages 292–305, 2023. 9

[12] Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image clustering with external guidance. In *ICML*, pages 27890–27902, 2024. 8

[13] Yinqi Li, Jiahe Zhao, Hong Chang, Ruibing Hou, Shiguang Shan, and Xilin Chen. un2clip: Improving clip's visual detail capturing ability via inverting unclip. In *NeurIPS*, 2025. 8

[14] 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. In *NeurIPS*, pages 2507–2521, 2022. 9

[15] Shijie Ma, Yuying Ge, Teng Wang, Yuxin Guo, Yixiao Ge, and Ying Shan. Genhancer: Imperfect generative models are secretly strong vision-centric enhancers. In *ICCV*, 2025. 7, 8, 9

[16] 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 *ICML*, pages 8748–8763, 2021. 7, 8

[17] 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. In *NeurIPS*, pages 87310–87356, 2024. 8

[18] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, pages 9568–9578, 2024. 8

[19] Wenxuan Wang, Quan Sun, Fan Zhang, Yepeng Tang, Jing Liu, and Xinlong Wang. Diffusion feedback helps clip see better. In *ICLR*, 2025. 7, 8

[20] 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 *ICLR*, 2024. 7

[21] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. 7

[22] Yuli Zhang, Huaiyu Wu, and Lei Cheng. Some new deformation formulas about variance and covariance. In *ICMIC*, pages 987–992, 2012. 2

[23] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 8