Reevaluating the Intra-Modal Misalignment Hypothesis in CLIP

Supplementary Material

A. On degrees of freedom

In Sec. 3.1 of the main paper, we presented a way to recover intra-modal similarities from inter-modal similarities. This solution assumed a set of text anchors for simplification. Here we show that even without this assumption, unique recovery of image-image similarities is possible:

Start with given inter-modal similarities

$$S_{inter} = X_T {X_I}^\top, \tag{4}$$

where the hidden underlying $X_T, X_I \in \mathbb{R}^{N\times d}$ have d-dim row vectors normalized to length 1 and $N \gg d$.

Decompose the $n \times n$ matrix S_{inter} via SVD:

$$S_{inter} = X_T {X_I}^\top = U\Sigma V^\top \tag{5}$$

where:

- $\Sigma \in \mathbb{R}^{d\times d}$ is a diagonal matrix, we do not further use it.
- $U, V \in \mathbb{R}^{N\times d}$ are orthogonal s.t. $U^\top U = I = V^\top V$.

Since the columns of V span the same space as the columns of (full-rank) X_I, we can write:

$$X_I = VC \tag{6}$$

for some $d \times d$ matrix C. Then:

$$S_{intra} = X_I X_I^\top = VCC^\top V^\top. \tag{7}$$

Let $Q = CC^\top$. Because of normalized X_I, we know:

$$\text{diag}(X_I X_I^\top) = \text{diag}(VQV^T) = \mathbf{1}_N. \tag{8}$$

This gives a linear system for the entries of $Q \in \mathbb{R}^{d\times d}$. Specifically, there are N quadratic forms

$$v_i^\top Q v_i = \sum_{j=1}^{d}\sum_{k=1}^{d} V_{ij}V_{ik}Q_{jk} = 1 \quad \text{for } i = 1, ..., N. \tag{9}$$

Solve by rearranging in a standard $Ax = b$ system,

$A:$ set the scalar product $V_{ij}V_{ik}$ as the entry of the ith row and $(jd - d + k)$th column in coefficient matrix $A \in \mathbb{R}^{N\times d^2}$.

$x:$ set flatten(Q) as the variable vector $x \in \mathbb{R}^{d^2}$.

$b:$ set $b = \mathbf{1}_N \in \mathbb{R}^N$.

There are $d(d+1)/2$ unknowns in x since Q is a symmetric $d \times d$ matrix. Since $N \gg d$, the linear system $Ax = b$ is overdetermined such that there is at most one solution for x and there are **no degrees of freedom**. In our recovery case, the solution exists; $b = 1_N$ is in the column space of A. Solving for x, then reshaping x back to Q, the intra-modal image-image similarities can be obtained via Eq. (7):

$$S_{intra} = VQV^T. \tag{10}$$

We provide a demonstration in PyTorch.

B. On the projection for more "classness"

In Sec. 3.3 of the main paper, we presented a way to reduce the semantics of an image embedding to its dominant concept by projecting onto axes spanned by class names.

Here we show that despite usage of text embeddings, this method, $PCA^{\leftarrow}$, still can be considered image-image comparison.

Interpretation - A neat way to illustrate this symmetry is by interpreting the projection of an image embedding $x_i \in \mathbb{R}^d$ as a sequence of three operations: a rotational change of basis ($Q^\top$) into the coordinate system defined by the principal components of class names, followed by a scaling (Λ) that preserves or cancels components, followed by the change back to the original basis (Q):

$$x_i^{\leftarrow} = Q\Lambda Q^\top x_i. \tag{11}$$

The resulting $x_i^{\leftarrow} \in \mathbb{R}^d$ is the projected image embedding.

The columns of $Q \in \mathbb{R}^{d\times d}$ contain the sorted eigenvectors (components) obtained by Eigendecomposition (PCA) of the covariance matrix of ImageNet class name text embeddings. The orthogonality of Q brings the isometric property that ensures Q and $Q^\top$ themselves preserve angles and distances, and hence also similarities.

The diagonal $\Lambda \in \mathbb{R}^{d\times d}$ determines the scaling of each basis vector. If $\Lambda = I$, then the projection has no effect ($x_i^{\leftarrow} = x_i$). We can instead set $\Lambda = \text{diag}(1..., 1, 0, ..., 0)$ to eliminate the components that do not explain much variance of class names. After (optional) re-projection to the original space via Q, the resulting $x_i^{\leftarrow}$ can be interpreted as the original x_i being preserved in selected directions, while cut off in the other.

Visualizations with UMAP, t-SNE and PCA in Fig. 7 demonstrate that $x_i^{\leftarrow}$ is indeed both globally and locally close to x_i. For visualization purposes, mean adjustment can be optionally performed via $\tilde{x_i}^{\leftarrow} = x_i^{\leftarrow} + \delta_\mu$,

$$\delta_\mu = Q(1 - \Lambda)Q^\top \mu_x, \tag{12}$$

i.e. adding back the part of the image embedding mean μ_x that was cut off during projection in Eq. (11). Since this is a translation, orthogonal to all $x_i^{\leftarrow}$, it has no effect on distances between $x_i^{\leftarrow}$; it only shifts them back towards the original center μ_x (compare Fig. 7 left and right). Besides these plots, Fig. 9 shows how CLIP-conditioned captions can still be generated with projected $x_i^{\leftarrow}$.

Concluding, it appears valid to interpret comparison of two $x_i^{\leftarrow}$ still as image-image comparison. Decent performance from this $PCA^{\leftarrow}$ comparison then suggests there is no issue with an intra-modal misalignment.

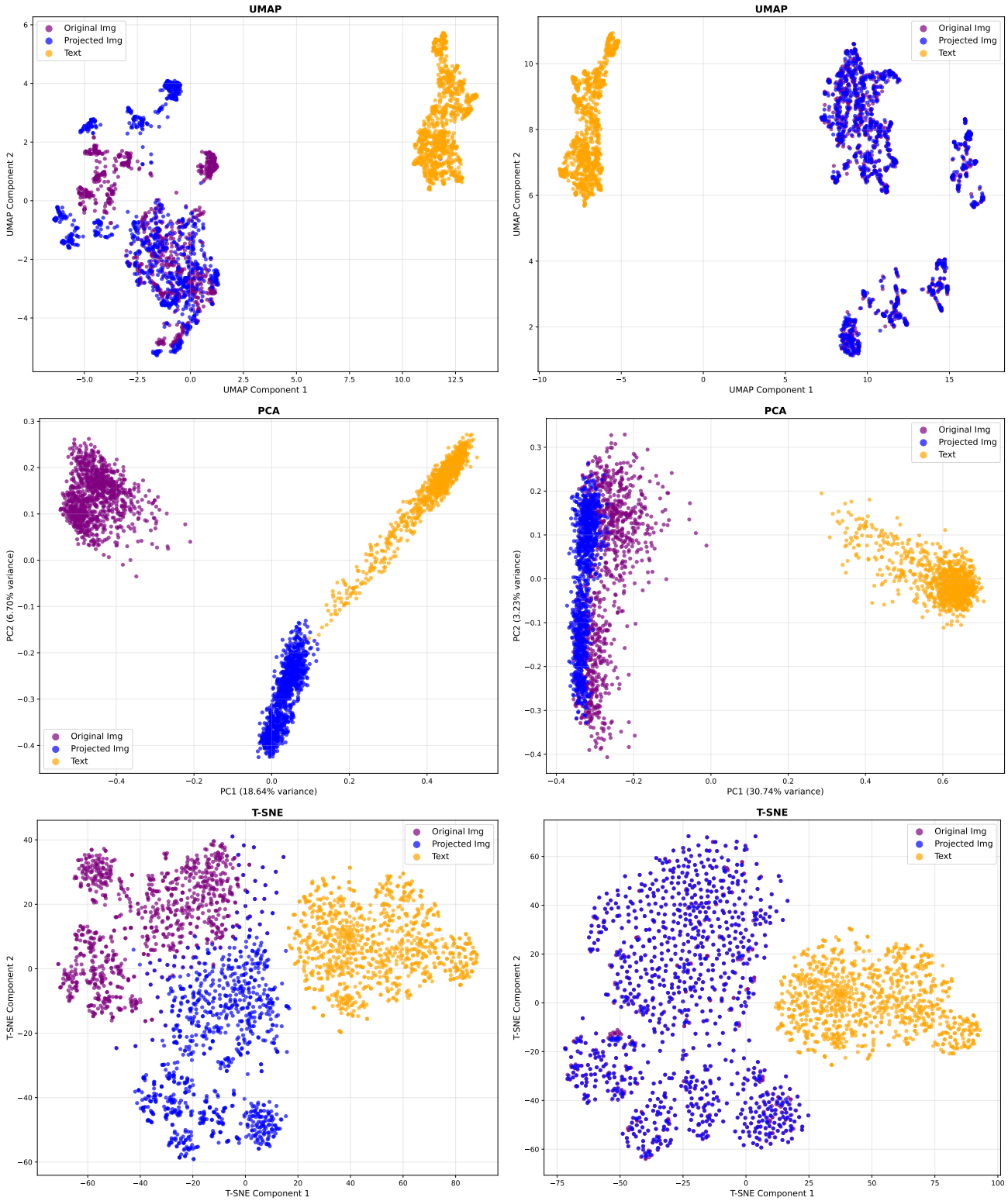

Figure 7. Distribution of CLIP embeddings: text (orange), original image (purple), projected image (blue). UMAP (top), PCA (middle), t-SNE (bottom). Left column: projection via Eq. (11). Right column: mean-adjusted projection via Eq. (12). *Interpretation*: (i) Like many previous studies noted, text embeddings and image embeddings lie in two different cones, such that we can find them clearly separated in the figure. We argued in the main paper this "modality gap" does not lead to an intra-modal misalignment. (ii) Projecting via $PCA^{\leftarrow}$ and re-projecting introduces a global shift (see left PCA) in $\mathbb{R}^{512}$ that is orthogonal to the principal components $\in \mathbb{R}^{256}$. (iii) For comparison in $\mathbb{R}^{512}$, we can compensate for this shift by adding back the mean of the canceled components (right column). ViT-B/16, ImageNet val.

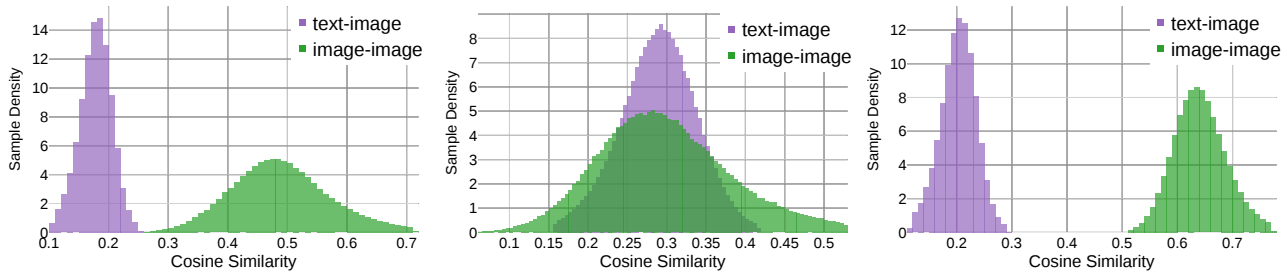

Figure 8. Modality gap: In the original CLIP space, image-image cosine similarities are high (left). With the projection from Eq. (11), these similarities seem to decrease (middle). Applying mean adjustment as in Eq. (12) removes this effect (right). *Interpretation*: Because mean adjustment preserves Euclidean distances, the observed changes arises solely from normalization: the projection zeros out components, reducing vector norms such that embeddings lie no more on, but inside the unit hypersphere. Normalization back on the hypersphere then squeezes the projected embeddings apart, leading to a lower cosine similarity value range (middle) compared to original CLIP (left). Adding back the canceled mean before normalization avoids this phenomenon (right). At the same time, there is no significant performance change between (middle) and (right), which once more illustrates that such cosine similarity histograms are insufficient indicators of alignment quality. ViT-B/16, ImageNet validation set.

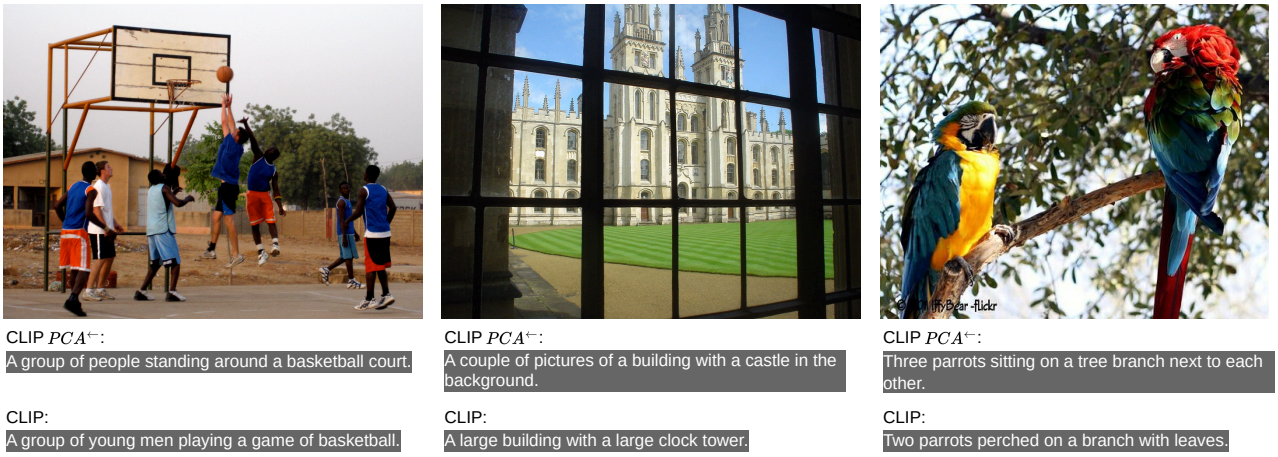

Figure 9. Captions generated with CLIP Prefix Captioning (Mokady *et al*. 2021), CLIP ViT-B/32, GPT-2. Swapping out the original CLIP image embedding x_i for our projected $x_i^{\leftarrow}$, we can observe the captions still cover the main objects. Samples are from datasets used for retrieval and few-shot classification in the main paper, left to right: SUN, ROxford, ImageNet.

Figure 10. The same experiment as in Fig. 9, but with the street scene dataset BDD100k used in the ablation in Sec. 4.3 of the main paper to validate our interpretation that the projection increases the "classness" of the image embedding, thereby being beneficial for classification-like tasks, but harmful for tasks such as daytime recognition because some information is lost. In line with the finding of the main paper, here we can see how the captions generated with $PCA^{\leftarrow}$ ignore that it is night (left, middle) and focus on the main objects only (right).