ALIGNMENT GOVERNS THE MODALITY GAP: UNDERSTANDING ITS INTERPLAY WITH UNIFORMITY AND DOWNSTREAM PERFORMANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastively trained vision-language models exhibit a persistent modality gap, with image and text representations occupying distinct regions of a shared embedding space, yet what governs this gap remains unclear. While one line of prior work has attributed the modality gap to modality-wise localization, we instead investigate the role of cross-modal alignment. Through an analysis of InfoNCE gradients, we show that InfoNCE does not explicitly promote alignment between the two modality regions. Using a controlled intervention that explicitly strengthens this form of alignment, we substantially reduce the modality gap, as measured by both positive-pair similarity and centroid distance, while largely preserving zero-shot classification and cross-modal retrieval performance. Notably, within-modality uniformity deteriorates as the gap decreases, demonstrating that improved uniformity is not necessary for gap reduction. We further uncover an empirical trilemma among positive-pair similarity, within-modality uniformity, and downstream performance: across controlled and alternative interventions, we find no examined configuration that simultaneously improves positive-pair similarity and uniformity while preserving downstream performance. Finally, we find that improved uniformity provides an additional route to reducing centroid distance without improving positive-pair similarity, yet targeting higher uniformity during training does not consistently yield a more favorable modality-gap–performance relationship under distribution shift. These findings identify direct cross-modal alignment as a mechanism governing modality-gap reduction, while revealing the metric-dependent role and limitations of within-modality uniformity.

1 INTRODUCTION

Contrastively trained vision-language models, such as CLIP (Radford et al., 2021), embed images and texts into a shared representation space and achieve strong performance across a wide range of downstream tasks (Jia et al., 2021; Gadre et al., 2023; Sun et al., 2023; Jose et al., 2025; Tschannen et al., 2025; Cao et al., 2026). Despite this success, image and text representations are known to occupy distinct regions of the shared space, forming a systematic separation known as the *modality gap* (Liang et al., 2022). This separation is commonly characterized by positive-pair similarity, which measures the cosine similarity between paired image and text representations (Goel et al., 2022; Jia et al., 2021), and centroid distance, which measures the distance between their modality centroids (Liang et al., 2022). While growing efforts have sought to understand and mitigate this gap, what governs its magnitude remains unclear.

One line of prior work attributes the modality gap to the localization of each modality within limited regions of the shared embedding space (Liang et al., 2022), motivating approaches that improve within-modality uniformity to spread the representations and increase their overlap (Fahim et al., 2024; Eslami & de Melo, 2025). Contrastive objectives such as InfoNCE (van den Oord et al., 2019), however, are known to shape representations through two complementary properties: alignment and uniformity (Wang & Isola, 2020). Alignment encourages positive pairs to be close, whereas uniformity encourages representations to spread uniformly over the representation hypersphere. A persistent modality gap implies that paired image and text representations remain separated, suggesting that reducing the gap should primarily be a matter of cross-modal alignment. This motivates

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

us to revisit the modality gap through the lens of alignment and ask whether directly strengthening alignment is sufficient to reduce the gap without improving uniformity.

To investigate this hypothesis, we analyze the feature gradients of InfoNCE and distinguish two notions of cross-modal alignment: *relative alignment*, which favors positive pairs over negative pairs, and *absolute alignment*, which directly brings paired image and text representations closer. Our analysis shows that the alignment mechanism of InfoNCE is inherently relative: its gradient is primarily composed of directions from negative samples toward the positive sample, without necessarily directing one representation toward its paired representation. Motivated by this observation, we introduce an intervention that explicitly strengthens absolute alignment. This absolute-alignment intervention substantially reduces the modality gap under both positive-pair similarity and centroid distance while largely preserving zero-shot classification and cross-modal retrieval performance. When absolute alignment is pushed further, however, downstream performance eventually degrades, consistent with prior observations that completely eliminating the modality gap can be harmful (Jiang et al., 2023). Importantly, within-modality uniformity consistently deteriorates as the modality gap decreases. These results demonstrate that improved uniformity is not necessary for modality-gap reduction and establish absolute alignment as a direct mechanism for reducing the gap.

We next ask whether the modality gap can be reduced while preserving both within-modality uniformity and downstream performance. To address this question, we extend the alignment intervention above with an explicit uniformity constraint and evaluate representation-level properties on the training distribution to enable controlled comparisons at fixed uniformity. We find an empirical trilemma among positive-pair similarity, within-modality uniformity, and downstream performance: at a comparable level of any one property, there is a trade-off between the other two. Specifically, at fixed uniformity, improving positive-pair similarity eventually degrades downstream performance; at comparable positive-pair similarity, improving uniformity degrades downstream performance; and, as observed above, at comparable downstream performance, higher positive-pair similarity is accompanied by deteriorated uniformity. Across additional interventions designed to decouple positive-pair similarity and within-modality uniformity, no examined configuration escapes this trilemma.

Finally, we ask whether the same empirical trilemma holds when the modality gap is measured by centroid distance. We find that it does not: improved uniformity can reduce centroid distance without improving positive-pair similarity, revealing a metric-dependent role of within-modality uniformity in modality-gap reduction. We then examine whether targeting higher uniformity during training yields a more favorable relationship between the modality gap and downstream performance under distribution shift. We find no consistent benefit: higher uniformity targets lead to a less favorable relationship for positive-pair similarity and provide no consistent advantage for centroid distance. These results question improved uniformity as a general strategy for modality-gap reduction, given its limited benefits and trade-offs with other desirable properties.

In summary, our contributions are threefold:

- **Cross-modal alignment provides a direct mechanism for modality-gap reduction.** Through gradient analysis of InfoNCE and a controlled intervention that strengthens absolute alignment, we show that both positive-pair similarity and centroid distance are improved while largely preserving zero-shot classification and cross-modal retrieval performance. Importantly, within-modality uniformity deteriorates as the gap is reduced, demonstrating that improved uniformity is not necessary for modality-gap reduction.
- **Positive-pair similarity exhibits an empirical trilemma with within-modality uniformity and downstream performance.** Through joint control of alignment and uniformity, we find that at a comparable level of any one property, there is a trade-off between the other two. Across alternative interventions designed to decouple positive-pair similarity and within-modality uniformity, no examined configuration escapes this trilemma.
- **Improved uniformity has limited utility as a general strategy for modality-gap reduction.** Uniformity can reduce centroid distance without improving positive-pair similarity, allowing centroid distance to escape the same empirical trilemma under controlled evaluation. However, uniformity is involved in the empirical trilemma with positive-pair similarity and downstream performance, and targeting higher uniformity during training provides no consistent benefit under distribution shift.

2 Related Work

Origins of the modality gap. Contrastively trained vision-language models learn image and text representations in a shared embedding space, yet the two modalities often occupy systematically separated regions, a phenomenon commonly referred to as the *modality gap* (Liang et al., 2022). Prior work has attributed this gap to several factors, including representation localization, optimization dynamics, cross-modal information asymmetry, and architectural inductive bias. The modality gap has been linked to image and text representations occupying limited regions of the embedding space, which can restrict their overlap; this localization has been observed in both initial and learned representations (Liang et al., 2022; Fahim et al., 2024; Eslami & de Melo, 2025). Optimization-based explanations include vanishing gap-reducing gradients (Chowers et al., 2026), slow convergence at low temperatures (Yaras et al., 2025), and modifications to the modality composition of negative samples (Oh et al., 2023; Yaras et al., 2025; Liu et al., 2026). Other work relates the modality gap to asymmetric information across modalities (Jiang et al., 2023; Schrodi et al., 2025). Architectural explanations instead point to dual encoders lacking an inductive bias toward mapping the modalities to a common region (Eslami & de Melo, 2025; You et al., 2022).

Uniformity and modality-gap reduction. Contrastive learning has been characterized by two properties, *alignment* and *uniformity*: alignment encourages positive pairs to be close, whereas uniformity encourages representations to spread uniformly over the representation hypersphere (Wang & Isola, 2020). Prior work has also suggested that a conflict between these properties can induce local minima with persistent modality separation (Shi et al., 2023). One line of work addresses the modality gap by promoting within-modality dispersion (Eslami & de Melo, 2025) or uniformity (Fahim et al., 2024) to alleviate modality-wise localization and increase cross-modal overlap. These studies establish an important connection between within-modality geometry and the modality gap, but leave open whether improved uniformity is necessary for gap reduction. We instead investigate whether directly strengthening cross-modal alignment can reduce the separation between modalities.

Why reduce the modality gap? Reducing the modality gap can benefit tasks requiring image and text representations to be treated symmetrically in a common embedding space, including SIMAT (Couairon et al., 2022; Oh et al., 2023; Fahim et al., 2024) and group-level semantic tasks (Grassucci et al., 2026). It has also been shown to improve robustness to representation-space perturbations in zero-shot classification and cross-modal retrieval (Chowers et al., 2026). However, its benefits for standard zero-shot classification and cross-modal retrieval performance are less clear, with prior work reporting marginal improvements or performance degradation as the gap is reduced (Schrodi et al., 2025; Chowers et al., 2026; Grassucci et al., 2026). Accordingly, we study modality-gap reduction not to improve zero-shot classification or cross-modal retrieval, but while seeking to preserve performance on these tasks.

3 From Relative to Absolute Alignment in InfoNCE

We investigate whether cross-modal alignment provides a direct mechanism for modality-gap reduction by analyzing the feature gradients of InfoNCE. Our analysis distinguishes two forms of cross-modal alignment: *relative alignment*, which favors positive pairs over negatives, and *absolute alignment*, which concerns reducing the separation between the two modality regions. We show that InfoNCE primarily induces relative alignment, then analyze when absolute alignment emerges from its gradient to identify the corresponding feature-update direction. Building on this analysis, we construct a controlled intervention that explicitly induces this direction.

3.1 Relative Alignment in InfoNCE

Let $\boldsymbol{f}_i^v$ and $\boldsymbol{f}_i^t$ denote the ℓ_2-normalized image and text representations of the i-th pair in a batch of size B. The image-to-text InfoNCE loss is

$$\mathcal{L}^{vt} = -\frac{1}{B}\sum_{i=1}^{B} \log p_{ii}^{vt}, \quad \text{and} \quad p_{ij}^{vt} = \frac{\exp(s_{ij}^{vt}/\tau)}{\sum_{k=1}^{B}\exp(s_{ik}^{vt}/\tau)}, \tag{1}$$

where $s_{ij}^{vt} = (\boldsymbol{f}_i^v)^T \boldsymbol{f}_j^t$ and τ is the temperature. The text-to-image loss $\mathcal{L}^{tv}$ is defined analogously with the modalities reversed, and symmetric InfoNCE is $\mathcal{L}_{\text{InfoNCE}} = (\mathcal{L}^{vt} + \mathcal{L}^{tv})/2$.

We focus on the gradient of $\mathcal{L}^{vt}$ with respect to the image representation $\boldsymbol{f}_i^v$; the corresponding text-side analysis follows symmetrically from $\mathcal{L}^{tv}$. Directional-loss ablations supporting this focus on the anchor-side gradient of each directional loss are provided in Appendix A.2. The resulting image-representation update direction is

$$-\nabla_{\boldsymbol{f}_i^v}\mathcal{L}^{vt} = \frac{1}{\tau B}\sum_{j\neq i} p_{ij}^{vt}\left(\boldsymbol{f}_i^t - \boldsymbol{f}_j^t\right). \tag{2}$$

This is a weighted combination of directions from negative text representations toward the positive text representation, thereby favoring the positive pair over negatives. We refer to this behavior as *relative alignment*. Relative alignment is particularly relevant to zero-shot classification and cross-modal retrieval, where predictions depend on relative similarities among candidate representations.

Relative alignment, however, does not specify how the separation between the two modality regions should be reduced. The update direction in Eq. (2) consists entirely of differences between text representations and therefore does not explicitly include a direction from the image representation toward the text modality region. This limitation is particularly relevant when the modalities initially occupy separated regions (Liang et al., 2022).

This geometric limitation is closely related to the recent analysis of Chowers et al. (2026). They show that contrastive learning can converge to a configuration in which the displacements between paired image and text representations share a common direction orthogonal to the affine subspaces of both modalities. Under their doubly-stochastic assumption, $-\nabla_{\boldsymbol{f}_i^v}\mathcal{L}^{tv}$ can likewise be expressed as a combination of text-representation differences (Appendix A.1). Hence, the full image-side gradient of symmetric InfoNCE is confined to their span. Since the modality-gap direction is orthogonal to this span, the gradient has no component reducing the modality gap. Our decomposition thus provides a complementary gradient-level interpretation of why such an orthogonal modality gap can persist.

3.2 Gradient Direction for Absolute Alignment

These observations raise a natural question: what feature-update direction should InfoNCE induce to directly reduce the separation between the two modality regions? Rather than specifying this direction heuristically, we identify it by analyzing when a gap-reducing update emerges from the InfoNCE gradient. We consider an idealized regime in which paired image and text representations are nearly aligned and negative samples are sufficiently dense.

Consider the hardest negative text $\boldsymbol{f}_{\hat{j}}^t$ for image $\boldsymbol{f}_i^v$, where $\hat{j} = \arg\max_{j\neq i} s_{ij}^{vt}$. The near alignment of paired representations gives $\boldsymbol{f}_{\hat{j}}^t \simeq \boldsymbol{f}_{\hat{j}}^v$. With sufficiently dense negative samples, we consider the regime in which the corresponding negative image lies close to the image anchor, such that $\boldsymbol{f}_{\hat{j}}^v \simeq \boldsymbol{f}_i^v$. Under these two approximations, the contribution of the hardest negative to the feature-update direction becomes

$$\boldsymbol{f}_i^t - \boldsymbol{f}_{\hat{j}}^t \simeq \boldsymbol{f}_i^t - \boldsymbol{f}_{\hat{j}}^v \simeq \boldsymbol{f}_i^t - \boldsymbol{f}_i^v. \tag{3}$$

Equation (3) identifies $\boldsymbol{f}_i^t - \boldsymbol{f}_i^v$ as the feature-update direction that emerges from InfoNCE in this idealized regime. Unlike the relative-alignment directions in Eq. (2), this direction directly connects the image representation to its paired text representation and thus reduces their cross-modal separation. We refer to this pair-directed behavior as *absolute alignment*.

3.3 A Controlled Intervention for Absolute Alignment

Based on the above analysis, we construct a controlled intervention. Our goal is not to improve downstream performance with a new contrastive objective, but to selectively strengthen absolute alignment while retaining the relative-alignment objective of standard InfoNCE. Accordingly, the intervention should (i) explicitly induce the absolute-alignment direction identified above and (ii) recover standard InfoNCE once a prescribed level of absolute alignment is reached.

Inducing the absolute-alignment direction. Equation (2) shows that each negative induces a feature-update direction from itself toward the positive representation. We exploit this property by introducing a positive-pair mixed negative $\tilde{\boldsymbol{f}}_i = \mathrm{sg}((\boldsymbol{f}_i^v + \boldsymbol{f}_i^t)/2)$ for image $\boldsymbol{f}_i^v$ and its paired

text $\boldsymbol{f}_i^t$. The additional negative $\tilde{\boldsymbol{f}}_i$ induces the feature-update direction $\boldsymbol{f}_i^t - \tilde{\boldsymbol{f}}_i = \frac{1}{2}(\boldsymbol{f}_i^t - \boldsymbol{f}_i^v)$, which is parallel to the absolute-alignment direction identified in Eq. (3). Importantly, stop-gradient must be applied to $\tilde{\boldsymbol{f}}_i$ to prevent gradients through the mixed representation, and $\tilde{\boldsymbol{f}}_i$ should not be normalized, as normalization would no longer guarantee the derived absolute-alignment direction. The text-side intervention is defined symmetrically.

Controlling the intervention. To control the strength of the absolute-alignment intervention, we introduce a margin m and modify the image-to-text probability as

$$\tilde{p}_{ij}^{vt} = \frac{\exp(s_{ij}^{vt}/\tau)}{\sum_{k=1}^{B} \exp(s_{ik}^{vt}/\tau) + \exp\left((\tilde{s}_i^v - m/2)/\tau\right)}, \tag{4}$$

where $\tilde{s}_i^v = (\boldsymbol{f}_i^v)^\top \tilde{\boldsymbol{f}}_i$. The text-to-image probability is modified symmetrically.

The contribution of the positive-pair mixed negative is progressively suppressed once its margin-adjusted similarity $\tilde{s}_i^v - m/2$ falls below the positive similarity s_{ii}^{vt}. Since $\tilde{s}_i^v = (1 + s_{ii}^{vt})/2$ for unit-normalized $\boldsymbol{f}_i^v$ and $\boldsymbol{f}_i^t$, this transition occurs at $s_{ii}^{vt} = 1 - m$. Thus, m controls the level of absolute alignment at which the optimization dynamics begin to approach standard InfoNCE.

We refer to the resulting objective as *Absolute Alignment InfoNCE (AA-InfoNCE)* and use it as a controlled intervention to investigate how strengthening absolute alignment affects the modality gap.

4 Experiments

We empirically investigate how absolute alignment and uniformity affect the modality gap and downstream performance. We first examine modality-gap reduction through absolute alignment, then investigate the empirical trilemma among positive-pair similarity, within-modality uniformity, and downstream performance. Finally, we assess the utility of improved uniformity as a general strategy for modality-gap reduction across gap metrics and under distribution shift.

4.1 Experimental Setup

Training. We conduct our main experiments on CC12M (Changpinyo et al., 2021), training OpenCLIP ViT-B/16 (Cherti et al., 2023) from scratch for 92k iterations with a batch size of 4,096. We additionally conduct the absolute-alignment experiments on LAION-400M (Schuhmann et al., 2021) to assess their generality, with results reported in Appendix D.3. Full training configurations are provided in Appendix B.

Controlled interventions. In Section 4.2, we use AA-InfoNCE alone and vary its margin m to control absolute alignment. In Sections 4.3 and 4.4, we additionally control within-modality uniformity using a regularizer that constrains it toward a prescribed target U_{target}:

$$\mathcal{L} = \mathcal{L}_{\text{AA-InfoNCE}} + \mathcal{L}_{\text{Uniformity}}, \tag{5}$$

$$\mathcal{L}_{\text{Uniformity}} = \frac{1}{2} \sum_{q \in \{v,t\}} \left\| U_q - U_{\text{target}} \right\|_1, \tag{6}$$

$$U_q = -\log \left(\frac{1}{B(B-1)} \sum_i \sum_{j \neq i} \exp\left(-2 \left\| \boldsymbol{f}_i^q - \boldsymbol{f}_j^q \right\|_2^2 \right) \right), \tag{7}$$

Here, $q \in \{v, t\}$ denotes the modality, and U_q is the negative of the standard uniformity metric (Wang & Isola, 2020) so that larger values indicate better uniformity.

Downstream performance. We evaluate downstream performance on ten image-classification datasets, including ImageNet-1K (Russakovsky et al., 2015), and two cross-modal retrieval datasets, COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014). For zero-shot classification, we use the dataset-specific prompt templates from the official CLIP implementation [1] and measure performance by top-1 accuracy. For cross-modal retrieval, we measure performance by the average of

[1] https://github.com/openai/CLIP/blob/main/data/prompts.md

image-to-text and text-to-image Recall@1. Since the number of datasets differs across task categories, we define overall downstream performance as the average of the mean classification and retrieval scores. Example dataset-wise comparisons between AA-InfoNCE and standard InfoNCE are provided in Appendix D.4.

Representation-level evaluation. We measure the modality gap using two complementary metrics: positive-pair similarity, the mean cosine similarity between paired image and text representations, and centroid distance, the Euclidean distance between the image and text feature centroids. Higher positive-pair similarity and lower centroid distance indicate a smaller modality gap. We report within-modality uniformity as $(U_v + U_t)/2$.

To better understand how the representation geometry relates to downstream performance, we additionally evaluate relative alignment, which measures how strongly positive pairs are favored over negatives. We quantify it using batch-wise positive-pair identification accuracy, averaged over the image-to-text and text-to-image directions.

We evaluate these representation-level properties on the CC12M training data to verify the intended effects of the controlled interventions and enable controlled comparisons. For the distribution-shift analysis, we additionally evaluate the CC12M-trained models on COCO and LAION-400M. Full evaluation details are provided in Appendix C.

4.2 Absolute Alignment Directly Reduces the Modality Gap

We first investigate whether strengthening absolute alignment alone can reduce the modality gap while preserving downstream performance. We vary the AA-InfoNCE margin m, where smaller m corresponds to a stronger absolute-alignment intervention.

Figures 1a and 1b show that strengthening absolute alignment reduces the modality gap under both metrics: as m decreases, positive-pair similarity increases and centroid distance decreases. Conversely, as the intervention weakens, both metrics approach those of standard InfoNCE, consistent with the intended behavior of AA-InfoNCE.

The two modality-gap metrics are related but not equivalent. Let $\boldsymbol{\delta}_i = \boldsymbol{f}_i^v - \boldsymbol{f}_i^t$ denote the displacement between paired image and text representations. The centroid distance satisfies

$$D_c = \left\| \frac{1}{B} \sum_i \boldsymbol{\delta}_i \right\|_2 \leq \frac{1}{B} \sum_i \|\boldsymbol{\delta}_i\|_2 = \frac{1}{B} \sum_i \sqrt{2\left(1 - (\boldsymbol{f}_i^v)^T \boldsymbol{f}_i^t\right)} \leq \sqrt{2(1 - S_{vt}^+)}, \tag{8}$$

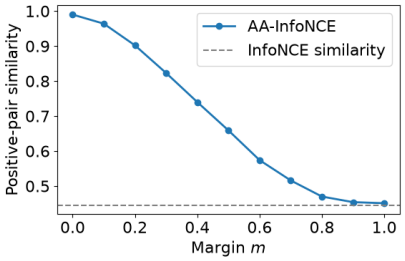
where S_{vt}^+ denotes the mean positive-pair similarity. Thus, increasing positive-pair similarity lowers the upper bound on centroid distance. The converse, however, does not hold: centroid distance becomes zero whenever $\sum_i \boldsymbol{\delta}_i = 0$, even if individual image–text pairs remain separated. This asymmetry leaves an additional route to reducing centroid distance without improving positive-pair similarity, which we examine in Section 4.4. A complementary decomposition in terms of within- and cross-modal unpaired similarities is provided in Appendix D.1.

We next examine downstream performance in Figure 1c. Downstream performance remains comparable to standard InfoNCE over a substantial range of positive-pair similarity. Pushing alignment further eventually degrades performance, consistent with prior findings that complete image–text alignment can be detrimental (Jiang et al., 2023). Thus, the modality gap can be substantially reduced through absolute alignment while largely preserving downstream performance.

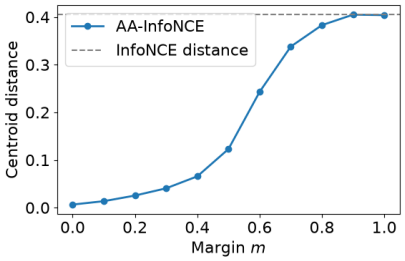
Finally, Figure 1d shows that within-modality uniformity consistently deteriorates as positive-pair similarity increases, demonstrating that improved uniformity is not necessary for modality-gap reduction. While these representation-level metrics are evaluated on the CC12M training distribution, we observe qualitatively consistent trends across different evaluation distributions and with models trained on LAION-400M (Appendices D.2 and D.3).

4.3 Positive-pair Similarity, Uniformity, and Downstream Performance

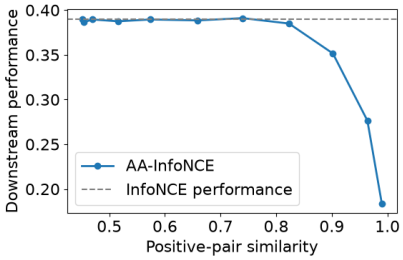
We next ask whether the modality gap and within-modality uniformity can be simultaneously improved without degrading downstream performance. To isolate the relationship between modality



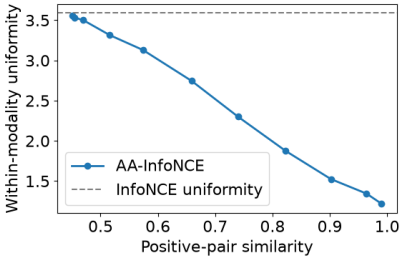
(a) Positive-pair similarity under varying AA-InfoNCE margins.



(b) Centroid distance under varying AA-InfoNCE margins.

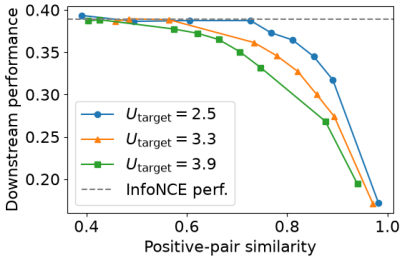


(c) Downstream performance under varying levels of positive-pair similarity.

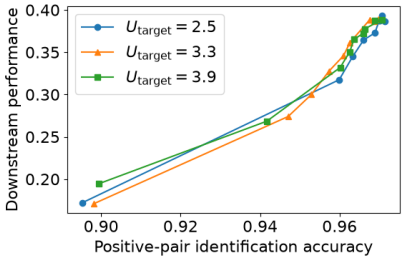


(d) Within-modality uniformity under varying levels of positive-pair similarity.

Figure 1: Effect of strengthening absolute alignment on the modality gap, downstream performance, and within-modality uniformity. Smaller AA-InfoNCE margins m correspond to stronger absolute alignment. Strengthening absolute alignment reduces both modality-gap measures while largely preserving downstream performance over a substantial range, despite deteriorating within-modality uniformity. Gray dashed lines indicate standard InfoNCE; all metrics approach their standard InfoNCE values as the intervention weakens, confirming the intended recovery behavior of AA-InfoNCE.



(a) Downstream performance across positive-pair similarity at fixed uniformity levels.



(b) Relationship between positive-pair identification accuracy and downstream performance.

Figure 2: Empirical trilemma among positive-pair similarity, within-modality uniformity, and downstream performance. (a) At fixed uniformity, increasing positive-pair similarity eventually degrades downstream performance, while at comparable positive-pair similarity, better uniformity is associated with lower downstream performance. (b) Higher positive-pair identification accuracy consistently corresponds to better downstream performance across different levels of absolute alignment and within-modality uniformity. Each curve connects models trained with different margins m at a fixed U_{target}.

gap and downstream performance at controlled levels of uniformity, we augment the alignment intervention with the uniformity constraint in Eq. (6). Since the uniformity constraint directly targets uniformity on the training distribution, we evaluate representation-level properties on the same distribution to enable controlled comparisons at fixed uniformity. The achieved uniformity closely matches each U_{target} across the margin sweep (Appendix E.1).

In this subsection, we analyze the modality gap in terms of positive-pair similarity. Figure 2a shows that downstream performance can be largely preserved over a range of positive-pair similarity at each uniformity level. However, as uniformity improves, downstream performance begins to degrade at

lower levels of positive-pair similarity. Thus, in our controlled experiments, positive-pair similarity and within-modality uniformity cannot be simultaneously improved while preserving downstream performance.

This behavior reflects two complementary trade-offs. At fixed uniformity, increasing positive-pair similarity eventually degrades downstream performance. Conversely, at comparable positive-pair similarity, better uniformity is associated with lower downstream performance. Together with our observation in Section 4.2 that higher positive-pair similarity is accompanied by deteriorated uniformity at comparable downstream performance, these results reveal an *empirical trilemma* among positive-pair similarity, within-modality uniformity, and downstream performance: at a comparable level of any one property, there is a trade-off between the other two.

We next examine whether the observed trilemma is specific to our controlled intervention. AlignCLIP (Eslami & de Melo, 2025) promotes within-modality dispersion to reduce modality-wise localization and reports improvements in both the modality gap and downstream performance, making it a potential counterexample to the observed trilemma. Under our evaluation, however, its publicly available model exhibits strong positive-pair similarity together with low within-modality uniformity (Appendix E.2). We additionally test several alternative interventions designed to decouple positive-pair similarity and within-modality uniformity (Appendix F). Across the configurations examined, we find no setting that escapes the empirical trilemma.

To better understand the downstream-performance component of the empirical trilemma, we examine relative alignment using positive-pair identification accuracy. Figure 2b shows that higher positive-pair identification accuracy consistently corresponds to better downstream performance across different levels of absolute alignment and within-modality uniformity. This observation provides a representation-level interpretation of the empirical trilemma in terms of absolute alignment, within-modality uniformity, and relative alignment.

4.4 The Utility and Limitations of Uniformity for Modality-Gap Reduction

We next examine whether the empirical trilemma identified for positive-pair similarity also holds when the modality gap is measured by centroid distance. Figure 3a shows that it does not. In particular, the $U_{\text{target}} = 2.5$ and 3.9 curves largely overlap, showing that substantially higher uniformity can coexist with similarly small centroid distance while preserving downstream performance.

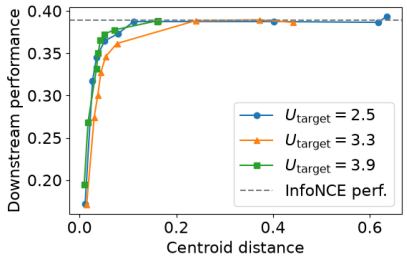
To understand this difference, Figure 3b relates centroid distance to positive-pair similarity across different uniformity levels. Better uniformity substantially reduces centroid distance at comparable positive-pair similarity, while having little effect on the range of positive-pair similarity. As discussed in Section 4.2, higher positive-pair similarity constrains centroid distance, but centroid distance can also be reduced without improving positive-pair similarity. The results here show that improved uniformity provides such an additional route, revealing its metric-dependent role in modality-gap reduction.

We have so far examined these relationships on the training distribution. We further examine whether targeting higher uniformity during training yields a more favorable relationship between the modality gap and downstream performance under distribution shift. Figure 4 evaluates the same CC12M-trained models on COCO. Even when the modality gap is measured by centroid distance, targeting higher uniformity provides no consistent benefit under distribution shift. Evaluation on LAION-400M likewise shows no consistent benefit from targeting higher uniformity for either modality-gap metric (Appendix G.1). These results further show that the benefit of targeting higher uniformity is limited under distribution shift.

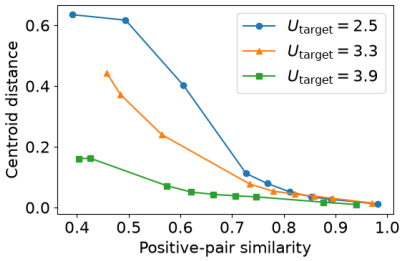
Together with the empirical trilemma involving positive-pair similarity and downstream performance, these findings question the utility of improving within-modality uniformity as a general strategy for modality-gap reduction.

5 Discussion

Our experiments do not establish the mechanism underlying the empirical trilemma, but we hypothesize that it may arise from the asymmetric and non-one-to-one correspondence between visual and textual information. Because multiple visually distinct images can correspond to similar tex-

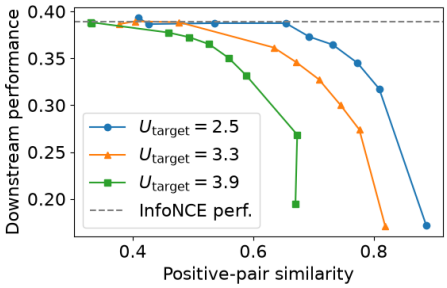


(a) Downstream performance across centroid distance at fixed uniformity levels.

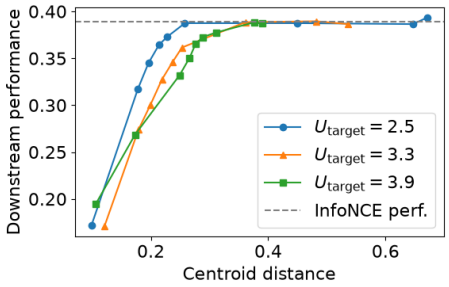


(b) Centroid distance across positive-pair similarity at different uniformity levels.

Figure 3: Metric-dependent role of within-modality uniformity in modality-gap reduction. (a) The $U_{\text{target}} = 2.5$ and 3.9 curves largely overlap, showing that substantially higher uniformity can coexist with similarly small centroid distance while preserving downstream performance. (b) Better uniformity reduces centroid distance at comparable positive-pair similarity, providing an additional route to centroid-distance reduction without improving positive-pair similarity.



(a) Relationship between downstream performance and positive-pair similarity.



(b) Relationship between downstream performance and centroid distance.

Figure 4: Modality-gap–performance relationships evaluated on COCO. The lowest uniformity target yields the most favorable relationship for both gap metrics.

tual semantics, strongly aligning each image with these shared semantics can indirectly draw the image representations closer to one another, deteriorating within-modality uniformity. Conversely, forcing such representations apart to improve uniformity may disrupt their cross-modal semantic relationships, potentially degrading relative alignment and downstream performance. Simultaneously enforcing strong absolute alignment and high uniformity may further encourage representations to specialize to the image–text correspondences observed during training, potentially limiting the cross-distribution utility of targeting higher uniformity.

6 Conclusion

We investigated what governs the modality gap in contrastively trained vision–language models and how its reduction interacts with within-modality uniformity and downstream performance. By explicitly strengthening absolute alignment, we substantially reduce the modality gap, as measured by both positive-pair similarity and centroid distance, while largely preserving downstream performance despite deteriorated within-modality uniformity. This demonstrates that improved uniformity is not necessary for modality-gap reduction.

Jointly controlling alignment and uniformity further reveals an empirical trilemma among positive-pair similarity, within-modality uniformity, and downstream performance. Although improved uniformity can provide an additional route to reducing centroid distance, targeting higher uniformity during training does not consistently yield a more favorable modality-gap–performance relationship under distribution shift. These results identify direct cross-modal alignment as a mechanism governing modality-gap reduction and question improved uniformity as a general strategy for reducing the gap.

AI USE STATEMENT

In this work, we used generative AI tools to help develop conceptual frameworks, provide critical ingredients for deriving mathematical claims, refine hypotheses, assist with translation, and interpret experimental results. We did not use generative AI tools to formulate mathematical claims, design or provide feedback on research methodology or experiments, or implement methods. Generating synthetic datasets, assisting in the writing of proofs, cleaning or reformatting datasets, and supporting qualitative or thematic data analysis were not applicable to this work. Additionally, we used generative AI tools to draft and edit parts of the manuscript, brainstorm ideas, suggest paper structure and titles, summarize and analyze existing literature, identify relevant literature, and search for information. We reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the details necessary to reproduce our experiments throughout the main paper and appendix. Section 3.3 describes the proposed AA-InfoNCE intervention, and Section 4.1 specifies the controlled interventions, evaluation metrics, and downstream evaluation protocol. Appendix B provides the full training configurations for the CC12M and LAION-400M experiments, and Appendix C describes the representation-level evaluation protocol.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 446–461, 2014.

Bingyi Cao, Koert Chen, Kevis-Kokitsi Maninis, Kaifeng Chen, Arjun Karpur, Ye Xia, Sahil Dua, Tanmaya Dabral, Guangxing Han, Bohyung Han, Joshua Ainslie, Alex Bewley, Mithun Jacob, René Wagner, Washington Ramos, Krzysztof Choromanski, Mojtaba Seyedhosseini, Howard Zhou, and Andre Araujo. Tipsv2: Advancing vision-language pretraining with enhanced patch-text alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29325–29335, 2026.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3558–3568, 2021.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023.

Rhea Chowers, Oshri Naparstek, Udi Barzelay, and Yair Weiss. Is the modality gap a bug or a feature? a robustness perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30288–30298, 2026.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Guillaume Couairon, Matthijs Douze, Matthieu Cord, and Holger Schwenk. Embedding arithmetic of multimodal queries for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4950–4958, 2022.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in CLIP. In *The Thirteenth International Conference on Learning Representations*, 2025.

Abrar Fahim, Alex Murphy, and Alona Fyshe. It's not a modality gap: Characterizing and addressing the contrastive gap. *arXiv preprint arXiv:2405.18570*, 2024.

Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 178–178, 2004.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. CyCLIP: Cyclic contrastive language-image pretraining. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Eleonora Grassucci, Giordano Cicchetti, Emanuele Frasca, Aurelio Uncini, and Danilo Comminiello. Closing the modality gap aligns group-wise semantics. In *The Fourteenth International Conference on Learning Representations*, 2026.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 4904–4916, 2021.

Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multimodal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7661–7671, 2023.

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. Dinov2 meets text: A unified framework for image- and pixel-level vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24905–24916, 2025.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. In *Second Workshop on Fine-Grained Visual Categorization*, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Hongyuan Liu, Qinli Yang, Wen Li, Zhong Zhang, Jiaming Liu, Wei Han, Zhili Qin, Jinxia Guo, and Junming Shao. The geometry of compromise: Unlocking generative capabilities via controllable modality alignment. *arXiv preprint arXiv:2604.00279*, 2026.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

Changdae Oh, Junhyuk So, Hoyoon Byun, YongTaek Lim, Minchul Shin, Jong-June Jeon, and Kyungwoo Song. Geodesic multi-modal mixup for robust fine-tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115:211–252, 2015.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Peiyang Shi, Michael C. Welle, Mårten Björkman, and Danica Kragic. Towards understanding the modality gap in CLIP. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2019.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9929–9939, 2020.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

Can Yaras, Siyi Chen, Peng Wang, and Qing Qu. Explaining and mitigating the modality gap in contrastive multimodal learning. In *The Second Conference on Parsimony and Learning (Proceedings Track)*, 2025.

Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 69–87, 2022.

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.

Table 1: Comparison of models trained with symmetric InfoNCE and its directional variant on CC12M and LAION-400M. In directional training, the image encoder is updated only by $\mathcal{L}^{vt}$ and the text encoder only by $\mathcal{L}^{tv}$. Despite removing the opposite-direction gradients, the directional variant retains substantial downstream performance and meaningful representation-level properties across both datasets.

Training dataset	Objective	Downstream performance	Positive-pair similarity	Centroid distance	U_v	U_t
CC12M	Symmetric	0.390	0.445	0.406	3.69	3.50
	Directional	0.329	0.392	0.201	3.75	3.81
LAION-400M	Symmetric	0.580	0.389	0.516	3.24	3.04
	Directional	0.512	0.383	0.396	3.40	3.50

A Additional Analysis of Directional InfoNCE Losses

We provide additional analysis of the directional InfoNCE losses discussed in Section 3.1. First, we show that, under the doubly stochastic assumption, the image-side gradient of $\mathcal{L}^{tv}$, like that of $\mathcal{L}^{vt}$, reduces to a weighted sum of directions from negative to positive text representations. We then compare directional training with standard symmetric InfoNCE to empirically assess how well the anchor-side directional gradients capture InfoNCE's representation-level behavior.

A.1 Gradient of the Opposite Directional Loss

For completeness, we consider the image-side gradient of the opposite directional loss $\mathcal{L}^{tv}$:

$$-\nabla_{\boldsymbol{f}_i^v}\mathcal{L}^{tv} = \frac{1}{\tau B}\sum_j \left(\delta_{ij} - p_{ji}^{tv}\right)\boldsymbol{f}_j^t, \tag{9}$$

where δ_{ij} denotes the Kronecker delta. Using $1 - p_{ii}^{tv} = \sum_{j\neq i} p_{ij}^{tv}$, Eq. (9) can be rewritten as

$$-\nabla_{\boldsymbol{f}_i^v}\mathcal{L}^{tv} = \frac{1}{\tau B}\sum_{j\neq i} p_{ji}^{tv}\left(\boldsymbol{f}_i^t - \boldsymbol{f}_j^t\right) + \frac{1}{\tau B}\left(\sum_j p_{ij}^{tv} - \sum_j p_{ji}^{tv}\right)\boldsymbol{f}_i^t. \tag{10}$$

The first term, as in $\mathcal{L}^{vt}$, is a weighted sum of directions from negative text representations toward the paired positive text representation. Under the doubly stochastic assumption,

$$\sum_j p_{ij}^{tv} = \sum_j p_{ji}^{tv} = 1, \qquad \forall i, \tag{11}$$

the second term vanishes, and $-\nabla_{\boldsymbol{f}_i^v}\mathcal{L}^{tv}$ also reduces to a weighted sum of text-feature differences.

A.2 Directional-Loss Ablation

The main-text analysis focuses on the gradient of each directional loss with respect to the representation that serves as its anchor. To assess whether this component alone can support meaningful representation learning, we train a directional variant in which the image encoder is updated only by $\mathcal{L}^{vt}$ and the text encoder only by $\mathcal{L}^{tv}$, stopping gradients from the opposite directional loss.

Table 1 compares models trained with the directional variant and standard symmetric InfoNCE on CC12M and LAION-400M, with representation-level properties evaluated on the respective training datasets. Despite removing the opposite-direction gradients, the directional variant retains substantial downstream performance and meaningful representation-level properties across both datasets. These results show that the anchor-side directional component analyzed in the main text alone captures substantial representation-learning behavior of InfoNCE, without implying that the opposite-direction gradients are negligible.

Table 2: Training configuration for the main experiments on CC12M.

Item	Value
Model architecture	OpenCLIP ViT-B/16
Training dataset	CC12M
Input resolution	224×224
Data augmentation	RandomResizedCrop, scale= $(0.9, 1.0)$, bicubic
Logit scale	Learnable $\log(1/\tau)$
Initial temperature τ	0.07
Minimum temperature τ	0.01
Training iterations	92k
Warmup iterations	10k
Batch size	4,096
Optimizer	AdamW
Optimizer hyperparameters	$\beta_1 = 0.9,\ \beta_2 = 0.98,\ \epsilon = 10^{-6}$
Learning rate	1×10^{-3}
Weight decay	0.2
Learning rate schedule	Cosine decay
Gradient clipping	5.0
Mixed precision	AMP (bfloat16)

Table 3: Training configuration for the additional experiments on LAION-400M.

Item	Value
Model architecture	OpenCLIP ViT-B/32
Training dataset	LAION-400M
Input resolution	224×224
Data augmentation	RandomResizedCrop, scale=$(0.9, 1.0)$, bicubic
Logit scale	Learnable $\log(1/\tau)$
Initial temperature τ	0.07
Minimum temperature τ	0.01
Training iterations	46k
Warmup iterations	10k
Batch size	65,536
Optimizer	AdamW
Optimizer hyperparameters	$\beta_1 = 0.9,\ \beta_2 = 0.98,\ \epsilon = 10^{-6}$
Learning rate	1×10^{-3}
Weight decay	0.2
Learning rate schedule	Cosine decay
Gradient clipping	5.0
Mixed precision	AMP (bfloat16)

B Training Details

We use the model implementations from OpenCLIP and train all models from scratch. Tables 2 and 3 summarize the training configurations for the CC12M and LAION-400M experiments, respectively.

C Representation-Level Evaluation Details

Modality gap and uniformity. For evaluation on CC12M and LAION-400M, we compute centroid distance and within-modality uniformity over batches of 65,536 image–text pairs and report the average over 16 iterations. Positive-pair similarity is computed as the mean cosine similarity of the paired representations over the same samples. For COCO, we select one caption per image, yielding 5k image–text pairs. The same selected captions are used across all evaluations.

Relative alignment. We evaluate positive-pair identification accuracy on the training dataset using the training batch size for each model and report the average over 16 iterations. Specifically, CC12M-trained models are evaluated on CC12M with a batch size of 4,096, whereas LAION-400M-trained models are evaluated on LAION-400M with a batch size of 65,536.

D Additional Results for Absolute Alignment

D.1 Centroid-Distance Decomposition and Similarity Structure

In Section 4.2, we showed that increasing positive-pair similarity lowers an upper bound on centroid distance. Here, we provide a complementary decomposition of centroid distance in terms of within- and cross-modal unpaired similarities and examine how this unpaired-similarity structure changes under the absolute-alignment intervention.

Let

$$\boldsymbol{\mu}_v = \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{f}_i^v, \qquad \boldsymbol{\mu}_t = \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{f}_i^t \tag{12}$$

denote the image and text centroids, respectively, such that the squared centroid distance is

$$D_c^2 = \|\boldsymbol{\mu}_v - \boldsymbol{\mu}_t\|_2^2. \tag{13}$$

We define the mean unpaired image–image, text–text, and image–text similarities as

$$S_{vv}^- = \frac{1}{B(B-1)}\sum_i \sum_{j\neq i} (\boldsymbol{f}_i^v)^\top \boldsymbol{f}_j^v, \tag{14}$$

$$S_{tt}^- = \frac{1}{B(B-1)}\sum_i \sum_{j\neq i} (\boldsymbol{f}_i^t)^\top \boldsymbol{f}_j^t, \tag{15}$$

$$S_{vt}^- = \frac{1}{B(B-1)}\sum_i \sum_{j\neq i} (\boldsymbol{f}_i^v)^\top \boldsymbol{f}_j^t, \tag{16}$$

and positive-pair similarity as

$$S_{vt}^+ = \frac{1}{B}\sum_i (\boldsymbol{f}_i^v)^\top \boldsymbol{f}_i^t. \tag{17}$$

Since all representations are ℓ_2-normalized, the centroid terms can be written as

$$\|\boldsymbol{\mu}_v\|_2^2 = \frac{1}{B} + \frac{B-1}{B} S_{vv}^-, \tag{18}$$

$$\|\boldsymbol{\mu}_t\|_2^2 = \frac{1}{B} + \frac{B-1}{B} S_{tt}^-, \tag{19}$$

$$\boldsymbol{\mu}_v^\top \boldsymbol{\mu}_t = \frac{1}{B} S_{vt}^+ + \frac{B-1}{B} S_{vt}^-. \tag{20}$$

Substituting these expressions into $D_c^2 = \|\boldsymbol{\mu}_v\|_2^2 + \|\boldsymbol{\mu}_t\|_2^2 - 2\boldsymbol{\mu}_v^\top \boldsymbol{\mu}_t$ gives

$$D_c^2 = \frac{B-1}{B}\left(S_{vv}^- + S_{tt}^- - 2S_{vt}^-\right) + \frac{2}{B}\left(1 - S_{vt}^+\right). \tag{21}$$

Thus, as B increases, the direct contribution of positive-pair similarity vanishes as $O(1/B)$, and centroid distance approaches

$$D_c^2 \simeq 2\left(\frac{S_{vv}^- + S_{tt}^-}{2} - S_{vt}^-\right). \tag{22}$$

Centroid distance is therefore increasingly determined by the mismatch between the mean within-modal similarity, $(S_{vv}^- + S_{tt}^-)/2$, and the cross-modal unpaired similarity S_{vt}^-.

Figure 5 examines these similarities for the CC12M-trained models from the absolute-alignment intervention experiments in Section 4.2. As the AA-InfoNCE margin m decreases and absolute alignment is strengthened, S_{vt}^- approaches the mean within-modal similarity $(S_{vv}^- + S_{tt}^-)/2$, reducing the similarity mismatch in Eq. (22). Thus, although our absolute-alignment intervention directly acts on paired representations, its reduction in centroid distance is accompanied by corresponding changes in the aggregate unpaired-similarity structure.

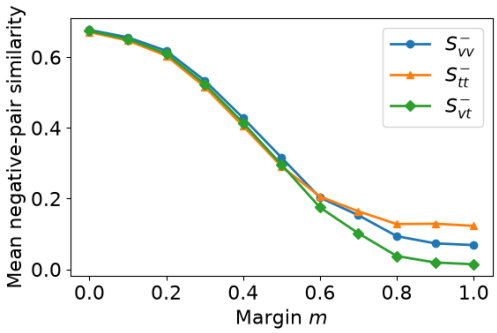


Figure 5: Mean unpaired similarities under varying AA-InfoNCE margins for CC12M-trained models evaluated on CC12M. As absolute alignment is strengthened (smaller m), S_{vt}^- approaches the mean within-modal similarity $(S_{vv}^- + S_{tt}^-)/2$, reducing the similarity mismatch that determines centroid distance in the large-batch regime.

D.2 Absolute-Alignment Intervention under Distribution Shift

We further evaluate the representation-level properties of the CC12M-trained models from Section 4.2 on LAION-400M and COCO. As shown in Figures 6 and 7, we observe qualitatively consistent trends across both evaluation distributions: strengthening the absolute-alignment intervention increases positive-pair similarity and reduces centroid distance while largely preserving downstream performance over a substantial range, despite deteriorating within-modality uniformity. Moreover, as the intervention weakens, all four properties approach their standard InfoNCE values, empirically confirming the intended recovery behavior of AA-InfoNCE.

D.3 Absolute alignment on LAION-400M

We further examine whether the findings of Section 4.2 generalize to models trained on LAION-400M. Figures 8, 9, and 10 evaluate the resulting models on LAION-400M, CC12M, and COCO, respectively. Across all three evaluation distributions, we observe qualitatively consistent trends: strengthening absolute alignment increases positive-pair similarity and reduces centroid distance while largely preserving downstream performance over a substantial range, despite deteriorating within-modality uniformity. Moreover, as the intervention weakens, all four properties approach their standard InfoNCE values.

On the LAION-400M training distribution, we additionally examine positive-pair identification accuracy. Figure 11 shows that higher positive-pair identification accuracy corresponds to better downstream performance across different levels of absolute alignment.

D.4 Dataset-wise downstream performance

To verify that the aggregate downstream score does not mask substantial improvements or degradations on individual datasets, we report dataset-wise results across ten image-classification datasets (Russakovsky et al., 2015; Krizhevsky, 2009; Fei-Fei et al., 2004; Krause et al., 2013; Nilsback & Zisserman, 2008; Bossard et al., 2014; Cimpoi et al., 2014; Parkhi et al., 2012; Cheng et al., 2017; Xiao et al., 2010) and two cross-modal retrieval datasets (Lin et al., 2014; Young et al., 2014) for AA-InfoNCE with $m = 0.4$. We select this setting because it achieves the strongest modality-gap reduction while largely preserving downstream performance for both the CC12M-trained and LAION-400M-trained models. As shown in Tables 4–7, we observe no substantial dataset-specific improvements or degradations relative to standard InfoNCE.

Table 4: Dataset-wise zero-shot classification performance of CC12M-trained models.

Loss	Avg.	ImageNet-1K	CIFAR100	Caltech-101	Stanford Cars	Oxford Flowers 102	Food-101	Describable Textures	Oxford-IIIT Pets	RESISC45	SUN397
InfoNCE	39.0	35.8	35.3	70.9	20.2	33.4	43.8	18.0	57.8	35.9	39.0
AA-InfoNCE	38.6	35.2	35.2	70.0	18.9	32.2	43.2	18.7	57.9	35.7	39.3

Table 5: Dataset-wise cross-modal retrieval performance of CC12M-trained models.

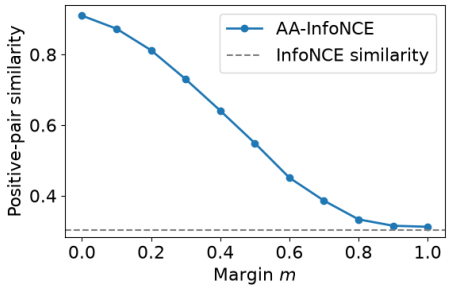
		COCO		Flickr30k	
Loss	Avg.	T2I R@1	I2T R@1	T2I R@1	I2T R@1
InfoNCE	38.9	22.9	34.0	41.9	56.8
AA-InfoNCE	39.6	22.2	33.4	43.5	59.5

Table 6: Dataset-wise zero-shot classification performance of LAION-400M-trained models.

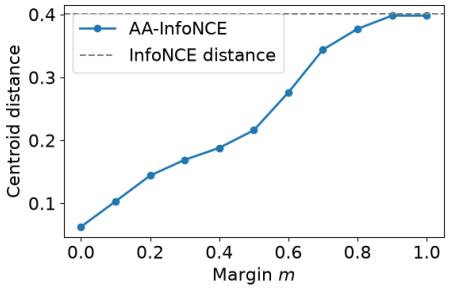
Loss	Avg.	ImageNet-1K	CIFAR100	Caltech-101	Stanford Cars	Oxford Flowers 102	Food-101	Describable Textures	Oxford-IIIT Pets	RESISC45	SUN397
InfoNCE	64.8	57.0	65.4	82.3	72.9	60.6	74.2	43.1	80.9	51.5	60.5
AA-InfoNCE	64.2	55.7	64.5	81.7	72.8	59.5	73.1	42.3	81.1	51.2	60.1

Table 7: Dataset-wise cross-modal retrieval performance of LAION-400M-trained models.

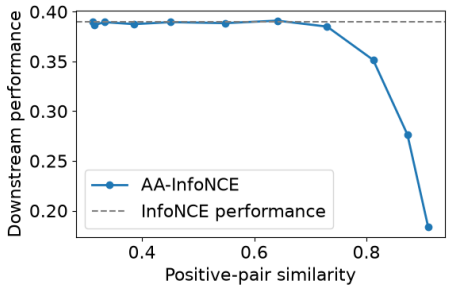
		COCO		Flickr30k	
Loss	Avg.	T2I R@1	I2T R@1	T2I R@1	I2T R@1
InfoNCE	51.1	30.5	46.8	55.6	71.5
AA-InfoNCE	51.1	29.7	46.9	54.6	73.2



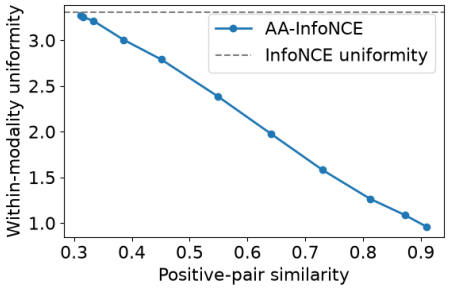
(a) Positive-pair similarity under varying AA-InfoNCE margins.



(b) Centroid distance under varying AA-InfoNCE margins.



(c) Downstream performance under varying levels of positive-pair similarity.



(d) Within-modality uniformity under varying levels of positive-pair similarity.

Figure 6: Absolute-alignment intervention for CC12M-trained models evaluated on LAION-400M. Strengthening absolute alignment reduces both modality-gap measures while largely preserving downstream performance over a substantial range, despite deteriorating within-modality uniformity. All metrics approach their standard InfoNCE values as the intervention weakens.

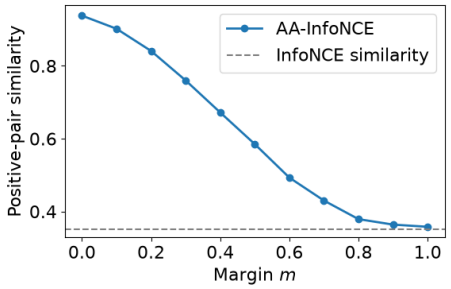
Table 8: Achieved image and text uniformity for models trained and evaluated on CC12M across different target uniformities and AA-InfoNCE margins.

$U_{\text{target}} = 2.5$			$U_{\text{target}} = 3.3$			$U_{\text{target}} = 3.9$		
m	U_v	U_t	m	U_v	U_t	m	U_v	U_t
0.0	2.47	2.50	0.0	3.28	3.30	0.0	3.87	3.88
0.2	2.46	2.50	0.2	3.27	3.30	0.2	3.89	3.88
0.25	2.46	2.50	0.25	3.27	3.30	0.4	3.89	3.90
0.3	2.46	2.50	0.3	3.26	3.30	0.45	3.89	3.90
0.35	2.47	2.50	0.35	3.26	3.30	0.5	3.89	3.90
0.4	2.47	2.50	0.4	3.27	3.30	0.55	3.89	3.90
0.6	2.49	2.50	0.6	3.28	3.30	0.6	3.89	3.90
0.8	2.49	2.50	0.8	3.29	3.30	0.8	3.89	3.90
1.0	2.49	2.50	1.0	3.29	3.30	1.0	3.89	3.90

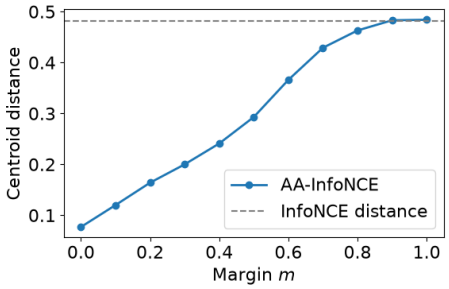
E Additional Results for the Empirical Trilemma

E.1 Achieved Uniformity under Target Constraints

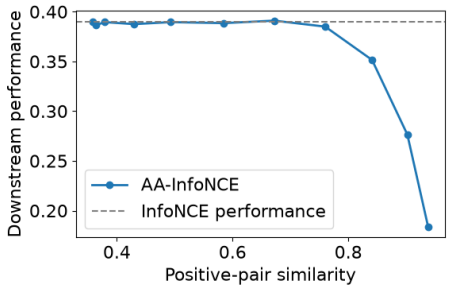
Table 8 reports the achieved image and text uniformity for models trained and evaluated on CC12M across different U_{target} values and AA-InfoNCE margins. Both U_v and U_t remain close to their prescribed targets and vary little across margins, validating the approximately fixed-uniformity comparisons in Section 4.3.



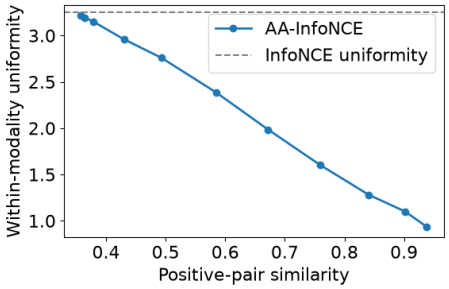
(a) Positive-pair similarity under varying AA-InfoNCE margins.



(b) Centroid distance under varying AA-InfoNCE margins.



(c) Downstream performance under varying levels of positive-pair similarity.



(d) Within-modality uniformity under varying levels of positive-pair similarity.

Figure 7: Absolute-alignment intervention for CC12M-trained models evaluated on COCO. Strengthening absolute alignment reduces both modality-gap measures while largely preserving downstream performance over a substantial range, despite deteriorating within-modality uniformity. All metrics approach their standard InfoNCE values as the intervention weakens.

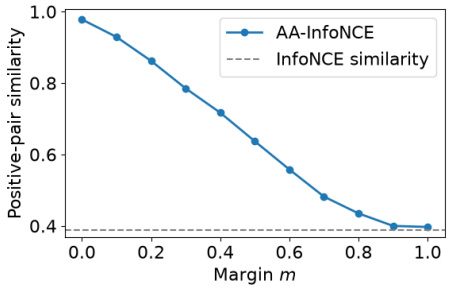
Table 9: Representation-level evaluation of AlignCLIP and AA-InfoNCE on COCO. AlignCLIP exhibits positive-pair similarity similar to that of AA-InfoNCE, while both models exhibit low within-modality uniformity.

Model	Positive-pair similarity	Centroid distance	U_v	U_t
AlignCLIP	0.668	0.370	2.08	1.62
AA-InfoNCE ($m = 0.4$)	0.672	0.241	1.94	2.03

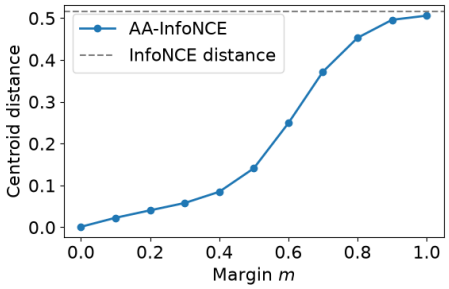
E.2 Evaluation of AlignCLIP

We further examine whether AlignCLIP (Eslami & de Melo, 2025), which promotes within-modality dispersion to increase cross-modal overlap, provides a counterexample to the empirical trilemma. We evaluate its publicly released model on COCO using our representation-level metrics. Table 9 reports the results together with AA-InfoNCE at $m = 0.4$, which largely preserves downstream performance, as a reference. AlignCLIP's positive-pair similarity is consistent with that reported in the original paper. Within-modality uniformity was not reported in the original study; under our metric, AlignCLIP yields $U_v = 2.08$ and $U_t = 1.62$.

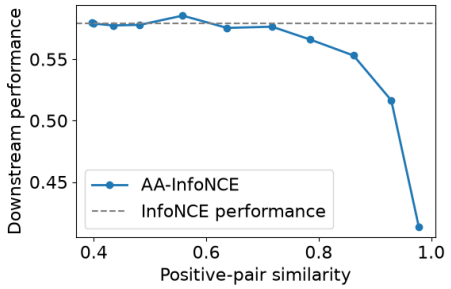
AlignCLIP is trained on CC12M with a ViT-B/16 backbone, as in our main experiments. Notably, its positive-pair similarity is similar to that of AA-InfoNCE, while both models exhibit low within-modality uniformity. Although differences in training configurations preclude a controlled comparison, the released AlignCLIP model therefore exhibits a representation geometry consistent with the trade-off observed in our experiments. Thus, we find no evidence that AlignCLIP provides a counterexample to the empirical trilemma.



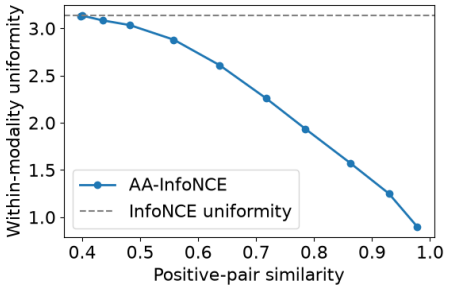
(a) Positive-pair similarity under varying AA-InfoNCE margins.



(b) Centroid distance under varying AA-InfoNCE margins.



(c) Downstream performance under varying levels of positive-pair similarity.



(d) Within-modality uniformity under varying levels of positive-pair similarity.

Figure 8: Absolute-alignment intervention for LAION-400M-trained models evaluated on LAION-400M. Strengthening absolute alignment reduces both modality-gap measures while largely preserving downstream performance over a substantial range, despite deteriorating within-modality uniformity. All metrics approach their standard InfoNCE values as the intervention weakens.

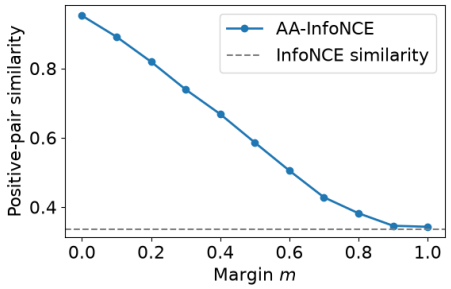
F Testing Alternative Explanations for the Empirical Trilemma

The empirical trilemma in Section 4.3 is established primarily through controlled interventions on absolute alignment and uniformity. We therefore examine whether the observed trade-offs could instead arise from specific limitations of these interventions. We consider three alternative explanations: (i) absolute alignment may over-align noisy or weakly matched positive pairs, thereby deteriorating within-modality uniformity, (ii) symmetrically enforcing uniformity across both modalities may be inappropriate given cross-modal information asymmetry, and (iii) the uniformity regularizer may disrupt local representation structure and thereby degrade relative alignment. For each hypothesis, we modify the corresponding intervention while preserving the overall experimental setting and examine whether the empirical trilemma can be avoided.

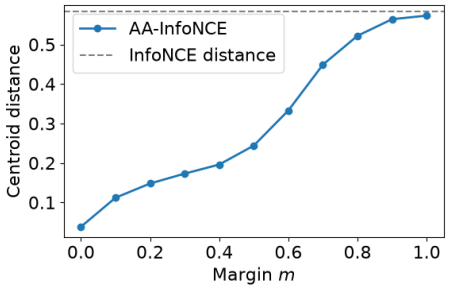
F.1 Over-Alignment of Noisy Positive Pairs

One possible explanation for the observed trade-off is that AA-InfoNCE over-aligns noisy or weakly matched positive pairs. Increasing the similarity of such inherently dissimilar pairs may increase similarities more broadly across the representation space, thereby deteriorating within-modality uniformity. To test this possibility, we restrict the absolute-alignment intervention to the 50% of positive pairs with the highest image–text similarity within each batch under the current model, while disabling the absolute-alignment intervention for the remaining pairs.

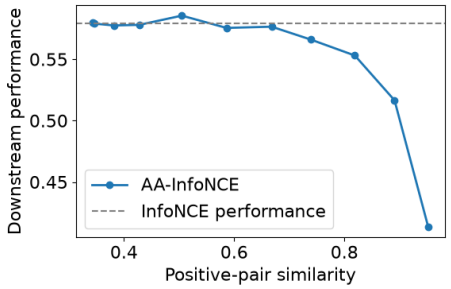
Figure 12a shows that this restriction produces nearly the same relationship between positive-pair similarity and downstream performance as the original intervention. The relationship between positive-pair similarity and within-modality uniformity is likewise nearly unchanged (Figure 12b). Thus, restricting absolute alignment to high-confidence positive pairs does not remove the observed trade-off between positive-pair similarity and within-modality uniformity, suggesting that the em-



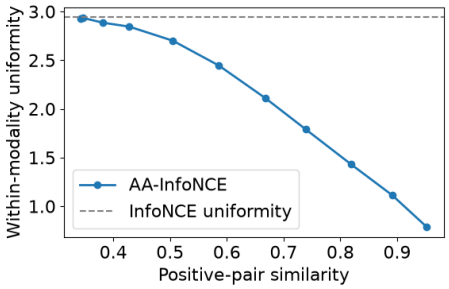
(a) Positive-pair similarity under varying AA-InfoNCE margins.



(b) Centroid distance under varying AA-InfoNCE margins.



(c) Downstream performance under varying levels of positive-pair similarity.



(d) Within-modality uniformity under varying levels of positive-pair similarity.

Figure 9: Absolute-alignment intervention for LAION-400M-trained models evaluated on CC12M. Strengthening absolute alignment reduces both modality-gap measures while largely preserving downstream performance over a substantial range, despite deteriorating within-modality uniformity. All metrics approach their standard InfoNCE values as the intervention weakens.

pirical trilemma cannot be explained solely by over-alignment of noisy or weakly matched positive pairs.

F.2 Symmetric Uniformity under Cross-Modal Information Asymmetry

A second possible explanation is that cross-modal information asymmetry makes it inappropriate to symmetrically enforce uniformity on both modalities in Eq. (6). To test this possibility, we vary the uniformity-loss formulation and examine whether the empirical trilemma can be avoided. We use AA-InfoNCE with $m = 0.4$ for all variants.

For the symmetric variant, we optimize

$$\mathcal{L} = \mathcal{L}_{\text{AA-InfoNCE}} - \frac{U_v + U_t}{2}, \tag{23}$$

where U_v and U_t are defined in Eq. (7). We compare this objective with image-only and text-only variants,

$$\mathcal{L} = \mathcal{L}_{\text{AA-InfoNCE}} - U_v, \qquad \mathcal{L} = \mathcal{L}_{\text{AA-InfoNCE}} - U_t, \tag{24}$$

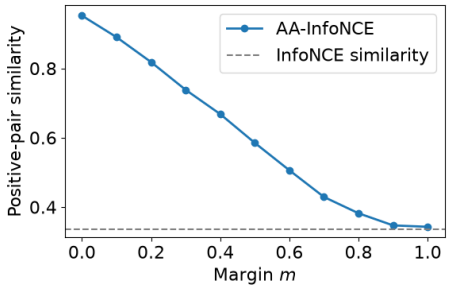
respectively. These variants test whether asymmetric uniformity regularization alleviates the trade-off.

We additionally evaluate Centroid Uniformity (Grassucci et al., 2026), which applies uniformity regularization to the centroid across modalities for each sample. In our two-modality setting, this centroid corresponds to the mean of the paired image and text representations:

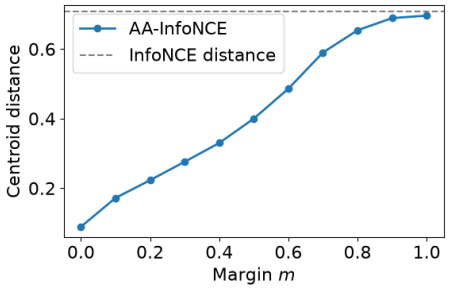
$$\boldsymbol{\mu}_i = \frac{\boldsymbol{f}_i^v + \boldsymbol{f}_i^t}{2}, \tag{25}$$

and optimize

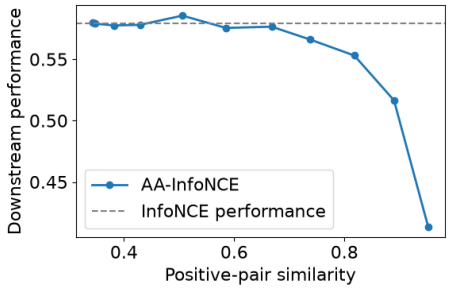
$$\mathcal{L} = \mathcal{L}_{\text{AA-InfoNCE}} - U_{\text{centroid}}, \tag{26}$$



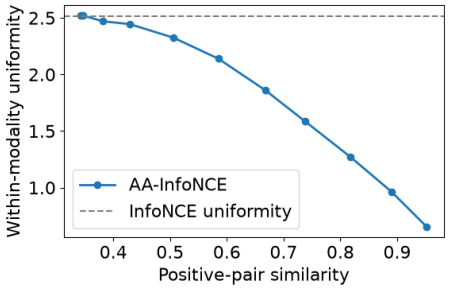
(a) Positive-pair similarity under varying AA-InfoNCE margins.



(b) Centroid distance under varying AA-InfoNCE margins.



(c) Downstream performance under varying levels of positive-pair similarity.



(d) Within-modality uniformity under varying levels of positive-pair similarity.

Figure 10: Absolute-alignment intervention for LAION-400M-trained models evaluated on COCO. Strengthening absolute alignment reduces both modality-gap measures while largely preserving downstream performance over a substantial range, despite deteriorating within-modality uniformity. All metrics approach their standard InfoNCE values as the intervention weakens.

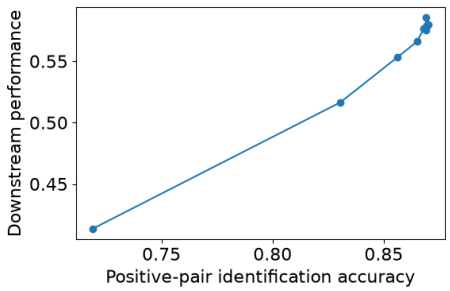
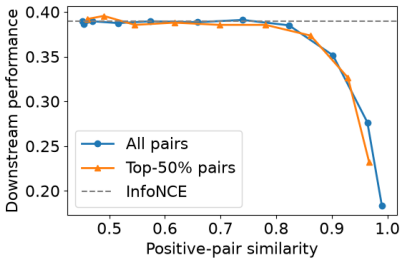


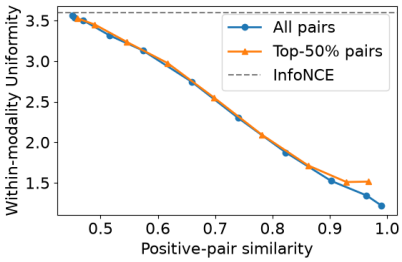
Figure 11: Relative alignment versus downstream performance for LAION-400M-trained models. Higher positive-pair identification accuracy corresponds to better downstream performance across different levels of absolute alignment, consistent with the results on CC12M.

where U_{centroid} is computed by applying Eq. (7) to the centroids $\{\boldsymbol{\mu}_i\}_{i=1}^{B}$.

Table 10 summarizes the results, with standard InfoNCE included as a reference. The choice of uniformity-loss formulation has only a modest effect on the resulting representations and downstream performance: all four variants achieve similarly high uniformity and positive-pair similarity, with substantially lower downstream performance than InfoNCE. Notably, even the symmetric and Centroid Uniformity variants, which achieve nearly identical within-modality uniformity, exhibit a trade-off between positive-pair similarity and downstream performance: the symmetric variant achieves higher positive-pair similarity but lower downstream performance. Thus, none of the alternative uniformity-loss formulations avoids the observed trade-off, suggesting that the empirical trilemma is not an artifact of symmetrically enforcing uniformity on the two modalities.



(a) Positive-pair similarity versus downstream performance.



(b) Positive-pair similarity versus within-modality uniformity.

Figure 12: Effect of restricting absolute alignment to high-confidence positive pairs. "Top-50% pairs" applies the absolute-alignment intervention only to the 50% of positive pairs with the highest image–text similarity, whereas "All pairs" applies it to all positive pairs. The relationships between positive-pair similarity and (a) downstream performance and (b) within-modality uniformity remain nearly unchanged, suggesting that over-alignment of noisy or weakly matched positive pairs alone cannot explain the observed trade-off.

Table 10: Ablation of uniformity-loss formulations with AA-InfoNCE ($m = 0.4$). Standard InfoNCE is included as a reference. Different uniformity formulations yield similar positive-pair similarity, within-modality uniformity, and downstream performance, and none avoids the observed trade-off.

Loss	Downstream performance	Positive-pair similarity	Centroid distance	U_v	U_t
Symmetric	0.331	0.745	0.029	3.91	3.92
Image only	0.335	0.738	0.059	3.91	3.87
Text only	0.335	0.726	0.101	3.83	3.92
Centroid	0.340	0.731	0.037	3.93	3.93
InfoNCE only	0.390	0.445	0.406	3.69	3.50

Table 11: Encoder-freezing ablation without uniformity regularization. The vision- and text-frozen variants are initialized from a model trained with standard InfoNCE and subsequently trained with AA-InfoNCE ($m = 0.4$) while freezing the corresponding encoder. Relative to AA-InfoNCE with the same margin, the frozen variants retain substantially higher within-modality uniformity but exhibit lower positive-pair similarity and downstream performance.

Model	Downstream performance	Positive-pair similarity	Centroid distance	U_v	U_t
Vision frozen	0.359	0.677	0.152	3.69	3.37
Text frozen	0.327	0.652	0.163	3.19	3.50
InfoNCE	0.390	0.445	0.406	3.69	3.50
AA-InfoNCE ($m = 0.4$)	0.391	0.740	0.066	2.26	2.34

F.3 Disruption of Local Representation Structure

A third possible explanation is that the uniformity regularizer disrupts local representation structure. Because the standard uniformity loss places greater weight on nearby representations, it may disproportionately separate semantically similar samples and disrupt local representation structure. Such disruption may degrade relative alignment and downstream performance, potentially accounting for the observed trade-off. We test this explanation using two interventions that avoid directly regularizing within-modality uniformity: one strengthens absolute alignment while freezing one encoder, and the other increases separation only among low-similarity negative pairs.

Encoder freezing. We first test whether the observed trade-off can be attributed to local structure distortion induced by the uniformity regularizer. Starting from a model trained with standard InfoNCE, we freeze either the vision or text encoder and continue training with AA-InfoNCE ($m = 0.4$), without any uniformity loss. The frozen modality therefore retains its InfoNCE uniformity, while the unfrozen modality remains free to adapt to the absolute-alignment objective.

Table 11 shows that, relative to standard InfoNCE, both frozen variants substantially improve positive-pair similarity, but at the cost of deteriorated uniformity in the unfrozen modality and lower downstream performance. Conversely, relative to AA-InfoNCE with the same margin $m = 0.4$, the frozen variants retain substantially higher within-modality uniformity but exhibit lower positive-pair similarity and downstream performance. Thus, even without a uniformity loss, preserving higher uniformity through encoder freezing does not escape the empirical trilemma.

Uniformity over dissimilar pairs. We next test whether the trade-off persists when uniformity is improved without directly separating semantically similar representations. In contrast to the standard uniformity regularizer, we construct a minimal intervention designed to improve uniformity indirectly by further separating only dissimilar within-modality pairs.

Let

$$s_{ij}^q = (\boldsymbol{f}_i^q)^\top \boldsymbol{f}_j^q, \qquad q \in \{v, t\}, \tag{27}$$

denote the within-modality similarity. For each anchor i, we define $\mathcal{D}_i^q$ as the 1% of samples within the batch with the lowest similarity to the anchor. We then add a uniformity margin m_U only to their similarities:

$$\tilde{s}_{ij}^q = s_{ij}^q + m_U \mathbb{I}[j \in \mathcal{D}_i^q]. \tag{28}$$

Using the resulting margin-adjusted similarities, we define an InfoNCE-style probability

$$p_{ij}^q = \frac{\exp(\tilde{s}_{ij}^q/\tau)}{\sum_{k=1}^B \exp(\tilde{s}_{ik}^q/\tau)}, \tag{29}$$

and define the corresponding within-modality loss as

$$\mathcal{L}_{\mathrm{DU}}^q = -\frac{1}{B}\sum_{i=1}^B \log p_{ii}^q. \tag{30}$$

Table 12: Uniformity over dissimilar pairs. AA-InfoNCE ($m = 0.4$) without $\mathcal{L}_{\mathrm{DU}}$ is included as a reference. Starting from AA-InfoNCE ($m = 0.4$), increasing the uniformity margin m_U improves within-modality uniformity while leaving positive-pair similarity nearly unchanged, but progressively degrades downstream performance.

Uniformity margin m_U	Downstream performance	Positive-pair similarity	Centroid distance	U_v	U_t
– (ref.)	0.391	0.740	0.066	2.26	2.34
0.6	0.388	0.740	0.063	2.28	2.36
0.8	0.379	0.740	0.068	2.32	2.40
1.0	0.372	0.741	0.077	2.37	2.46

Since the representations are normalized, the self-similarity is $s^q_{ii} = 1$. Adding m_U therefore makes the selected dissimilar samples harder negatives, encouraging their original similarities to decrease further. We apply this loss symmetrically to the image and text modalities:

$$\mathcal{L} = \mathcal{L}_{\text{AA-InfoNCE}} + \mathcal{L}_{\mathrm{DU}}, \tag{31}$$

$$\mathcal{L}_{\mathrm{DU}} = \frac{1}{2}\left(\mathcal{L}^v_{\mathrm{DU}} + \mathcal{L}^t_{\mathrm{DU}}\right). \tag{32}$$

Table 12 shows that increasing m_U progressively improves within-modality uniformity while leaving positive-pair similarity nearly unchanged. Under this nearly constant positive-pair similarity, downstream performance consistently decreases as uniformity improves. Thus, the trade-off persists even when uniformity is improved indirectly by separating only dissimilar within-modality pairs, without directly separating local neighbors. This suggests that the empirical trilemma cannot be explained solely by disruption of local representation structure induced by the uniformity regularizer.

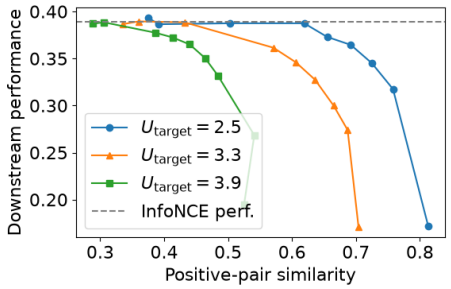
Across these experiments, the empirical trilemma persists after modifying both the absolute-alignment and uniformity interventions. Restricting absolute alignment to high-confidence positive pairs, replacing the symmetric uniformity loss with asymmetric or centroid-based alternatives, removing uniformity regularization through encoder freezing, and indirectly improving uniformity by separating only dissimilar pairs all fail to escape the observed trade-off. These results do not establish that the trilemma is unavoidable, but suggest that it cannot be readily attributed to the specific limitations of the controlled interventions used in the main experiments.

G Additional Results under Distribution Shift

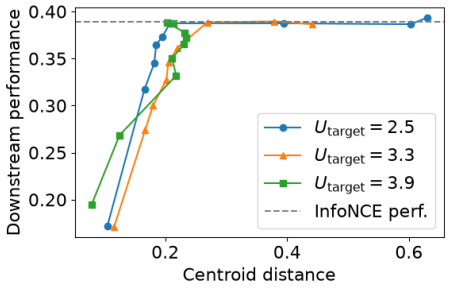
We provide additional analyses of the CC12M-trained models under distribution shift. We first evaluate the modality-gap–performance relationship on LAION-400M, complementing the COCO results in Section 4.4, and then examine the degradation of the modality-gap metrics across evaluation distributions.

G.1 Modality-Gap–Performance Relationship on LAION-400M

We evaluate the CC12M-trained models from Section 4.4 on LAION-400M to further assess whether targeting higher uniformity during training yields a more favorable relationship between the modality gap and downstream performance under distribution shift. Figure 13 shows no consistent benefit from targeting higher uniformity. For positive-pair similarity, higher uniformity targets yield a less favorable relationship with downstream performance. For centroid distance, the relationship varies non-monotonically across uniformity targets, with no consistent advantage from targeting higher uniformity. These results are consistent with the COCO evaluation in showing no benefit from targeting higher uniformity.



(a) Relationship between downstream performance and positive-pair similarity.

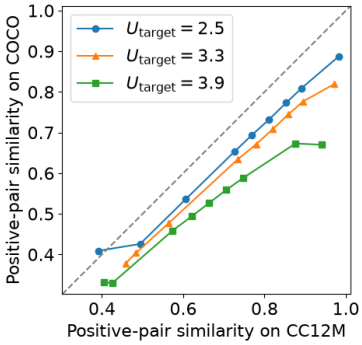


(b) Relationship between downstream performance and centroid distance.

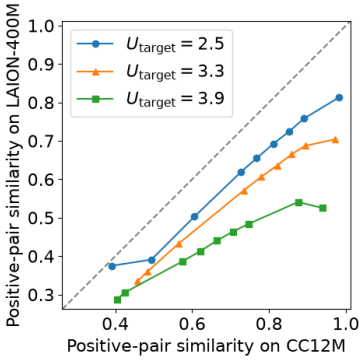
Figure 13: Modality-gap–performance relationships for CC12M-trained models evaluated on LAION-400M. Lower uniformity targets yield a more favorable relationship for positive-pair similarity, while higher targets provide no clear advantage for centroid distance.

G.2 Cross-Distribution Degradation of Modality-Gap Metrics

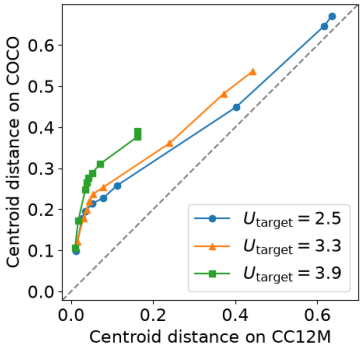
As a complementary analysis, we examine how positive-pair similarity and centroid distance change under distribution shift from CC12M to COCO and LAION-400M across models trained with different uniformity targets. Figure 14 shows the results for both evaluation distributions. For positive-pair similarity, models trained with higher uniformity targets exhibit greater degradation across both evaluation distributions at comparable similarity on CC12M. For centroid distance, higher uniformity targets are associated with greater degradation on COCO, whereas no clear uniformity-dependent trend is observed on LAION-400M.



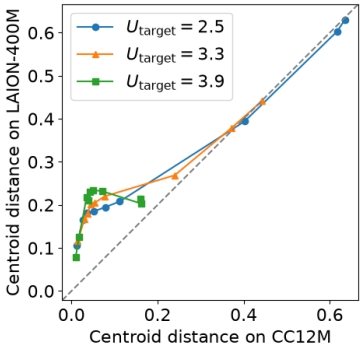
(a) Positive-pair similarity on COCO.



(b) Positive-pair similarity on LAION-400M.



(c) Centroid distance on COCO.



(d) Centroid distance on LAION-400M.

Figure 14: Cross-distribution degradation of modality-gap metrics for CC12M-trained models. Gray dashed lines indicate $y = x$. Higher uniformity targets are associated with greater degradation in positive-pair similarity on both COCO and LAION-400M. For centroid distance, greater degradation is observed with higher uniformity targets on COCO, while no clear uniformity-dependent trend is observed on LAION-400M.