ON THE ORIGIN OF THE MODALITY GAP: A MECHANISTIC PERSPECTIVE VIA CORRESPONDENCE NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive vision-language models exhibit a persistent modality gap despite being trained to align images and text, yet its underlying mechanism remains unclear. We provide a mechanistic account demonstrating that a small number of high-offset dimensions disproportionately shift the ground truth retrieval rank of samples with weak, ambiguous, or erroneous image–text correspondence through their cross-modal interaction. Controlled perturbations of correspondence quality reproduce this behavior, providing causal evidence that correspondence noise contributes to the emergence and function of the gap. We introduce a correspondence noise score, show consistency across checkpoints and diverse pretrained models, and demonstrate that filtering correspondence noise before training jointly reduces the modality gap and improves downstream performance. Finally, we show that gap magnitude is positively but inconsistently associated with performance across converged models, making it an unreliable proxy for representation quality, while its functional importance is primarily geometric rather than semantic.

1 INTRODUCTION

Contrastive vision–language models such as CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) have become a central paradigm for learning transferable multimodal representations by embedding images and text into a shared space. These models support strong zero-shot transfer and have become the foundation of many contemporary multimodal systems. Yet, despite being explicitly trained for cross-modal alignment, image and text representations remain systematically separated in the representation space (Liang et al., 2022). Commonly referred to as the modality gap, this phenomenon has been observed across architectures, datasets, training regimes and modalities.

Existing explanations address different aspects of why such a separation may arise. Theoretical and optimization-based accounts relate the gap to properties of the contrastive objective, initialization, representation geometry, or asymmetries between the two encoders (Liang et al., 2022; Shi et al., 2023; Fahim et al., 2024; Yaras et al., 2025; Mager et al., 2026). More recent accounts emphasize structural differences between modalities: images typically contain substantially more information than is captured by their paired captions, such that perfect alignment may be neither possible nor desirable (Schrodi et al., 2025; Jiang et al., 2023; Ramasinghe et al., 2024). These explanations are supported by theoretical arguments and empirical evidence, but emphasize different and, in some cases, competing factors underlying the modality gap. Much of the existing evidence establishes conditions under which a gap can arise or be modulated, but does not yet identify the mechanism underlying the gap in trained models or clarify its functional role in the learned representation.

This incomplete understanding is consequential because an increasing body of work explicitly attempts to close or reduce the modality gap (Zhang et al., 2023; Oh et al., 2023; Eslami & de Melo, 2025; Mistretta et al., 2025; Yamaguchi et al., 2025; An et al., 2025; Wang et al., 2026b). Yet, the empirical motivation for doing so remains unclear: changes in the magnitude of the gap do not consistently translate into corresponding changes in downstream performance (Schrodi et al., 2025; Chowers et al., 2026; Grassucci et al., 2026). Consequently, a smaller gap cannot simply be assumed to represent a better aligned or more useful representation. Before intervening on the gap, it is therefore important to understand its functional role.

We find that the answer is tied to imperfect image–text correspondence (i.e., correspondence noise). The modality gap is concentrated in a small number of dimensions in which the mean activation of

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.

one modality is strongly shifted away from zero (i.e., high offset) while the other remains close to zero. These dimensions do more than globally separate image and text representations: they exert a disproportionate influence on cross-modal similarity rankings. Crucially, this influence is highly sample-specific. Removing a gap dimension leaves most pairs almost unchanged while sharply shifting the rank of the paired ground truth (GT) match for a small subset of samples, which are systematically characterized by weak, ambiguous, or erroneous image–text correspondence. The effect follows directly from the cross-modal similarity computation: for a fixed query, the large offset in one modality multiplies candidate-specific values in the other, amplifying differences between the paired match and competing candidates. The resulting re-ranking therefore depends on the interaction between both modalities rather than on the offset alone. This suggests that the modality gap serves a sample-specific role in reshaping similarity rankings when image–text correspondence is weak or noisy. Our contributions are as follows:

- **A mechanistic account of the modality gap.** We show that sparse, high-offset dimensions exert disproportionate and sample-specific influence on cross-modal similarity rankings, with the strongest effects concentrated on weak or noisy image–text correspondences (Section 4.1). Controlled perturbations of correspondence quality reproduce this behavior, providing evidence that correspondence noise is a factor in the emergence and function of the modality gap (Section 4.2)
- **A sample-level measure of correspondence noise.** To quantify this mechanism, we introduce the *CN score* (i.e., correspondence noise score), which measures how strongly an individual image–text pair depends on modality-gap dimensions (Section 3.1). We show that the *CN score* is substantially consistent across intermediate training checkpoints and across diverse pretrained models (Section 4.3), and that filtering correspondence noise before training jointly reduces the modality gap and improves downstream performance (Section 4.4).
- **A reassessment of what the modality gap means for representation quality**. Across converged models, gap magnitude shows task-dependent positive associations with downstream performance, demonstrating that a smaller gap is not in itself indicative of a better representation. We further show that the apparent importance of gap dimensions (Schrodi et al., 2025) largely stems from the global geometry they impose on the representation, rather than from unusually rich sample-specific semantic information (Section 4.5).

2 Related Work

Modality Gap. Following the initial characterization of the modality gap by Liang et al. (2022), several explanations have been proposed, including properties of contrastive optimization, representation geometry, and encoder asymmetries (Shi et al., 2023; Fahim et al., 2024; Yaras et al., 2025; Mager et al., 2026), as well as information imbalance between modalities (Schrodi et al., 2025). These accounts have motivated methods that reduce or compensate for cross-modal separation (Zhang et al., 2023; Oh et al., 2023; Eslami & de Melo, 2025; Mistretta et al., 2025; Yamaguchi et al., 2025; An et al., 2025; Wang et al., 2026b; Li et al., 2026), while other work questions whether closing the gap is consistently beneficial (Jiang et al., 2023; Ramasinghe et al., 2024; Huang et al., 2025; Chowers et al., 2026; Grassucci et al., 2026). Several findings also suggest a relationship between pair-wise correspondence and the modality gap. Liang et al. (2022) show in a small simulation that removing geometrically mismatched pairs eliminates a temperature-dependent repulsive effect; Schrodi et al. (2025) find that reducing caption information increases the gap; and Yaras et al. (2025) show that mismatched pairs can enlarge it during training. Conversely, Wang et al. (2026a) study the gap as a source of error under noisy correspondence. These findings point toward a relationship between correspondence and the modality gap, but do not explain how correspondence quality relates to the sparse gap dimensions or why their effect is strongly sample-specific.

Correspondence Noise. Noisy-correspondence learning studies partially mismatched or unreliable paired observations (Huang et al., 2021; Yang et al., 2021), with methods based on loss statistics, cross-modal consistency, neighborhood structure, and geometric relationships (Yang et al., 2023; Qin et al., 2023; Dang et al., 2024; Yang et al., 2024; Zhao et al., 2024). Such noise is common in web-scale vision–language data (Jia et al., 2021; Gadre et al., 2023; Fang et al., 2024; Xu et al., 2024). Whereas this literature primarily treats correspondence noise as a data-quality problem to detect or mitigate, we study its relationship to the emergence and function of the modality gap.

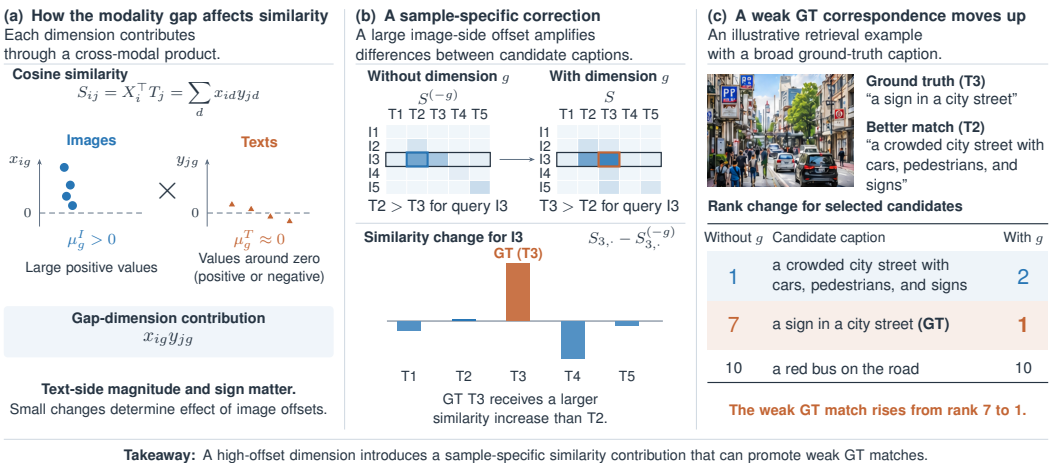

Figure 1: Mechanism of sample-specific re-ranking by modality-gap dimensions. (Examples of pairs of weak GT correspondence are illustrated in Figure 4).

3 Measuring Correspondence Noise and the Modality Gap

3.1 Correspondence Noise Score

We quantify correspondence noise through the extent to which the GT retrieval rank of an image–text pair is affected by modality-gap dimensions. Since GT retrieval rank depends on the candidate images or texts in the retrieval set (i.e., competitors), the score is relative to a given model and evaluation dataset rather than an intrinsic property of a sample.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote a dataset of paired images and texts, and let $X_i, T_i \in \mathbb{R}^D$ be their normalized representations under a given model. We compute correspondence scores within batches $B \subset \{1, \ldots, N\}$ of size M. For image-to-text retrieval, the similarity between image i and text j within a batch is $S_{ij} = X_i^\top T_j$, with $i, j \in B$.

For image query i, its paired text $j = i$ constitutes the GT. We define its rank as

$$r_i^{I \to T}(B) = 1 + \sum_{j \in B} \mathbf{1}[S_{ij} > S_{ii}]. \tag{1}$$

We first define the effect of an arbitrary representation dimension d. Let $S_{ij}^{(-d)}$ denote the similarity between image i and text j after removing dimension d from both representations and renormalizing the remaining coordinates. Recomputing the GT rank under $S^{(-d)}$ gives $r_i^{(-d, I \to T)}(B)$. We define the dimension-specific absolute GT rank change as

$$c_{i,d}^{I \to T}(B) = \left| r_i^{(-d, I \to T)}(B) - r_i^{I \to T}(B) \right|. \tag{2}$$

The quantity $c_{i,d}$ therefore measures the change in the retrieval position of the paired GT induced by removing dimension d.

Let $\mathcal{G} = \{g_1, \ldots, g_K\}$ denote the modality-gap dimensions, identified heuristically by pronounced one-sided mean offsets (typically $\gtrsim 0.2$) at intermediate or final checkpoints. We define the image-to-text *CN score* (i.e., correspondence noise score) as the mean absolute GT rank change across these dimensions, and average both retrieval directions for the final score:

$$CN_i^{I \to T}(B) = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} c_{i,g}^{I \to T}(B), \qquad CN_i(B) = \frac{1}{2} \left(CN_i^{I \to T}(B) + CN_i^{T \to I}(B) \right). \tag{3}$$

Here, $CN_i^{T \to I}(B)$ is defined analogously for text-to-image retrieval.

Because the rank of a GT pair depends on the particular set of competing samples in its batch, a single batch provides only one estimate of its correspondence score. In practice, we therefore

evaluate each sample across repeated stratified random batch assignments and average the resulting scores,

$$\widehat{CN}_i = \frac{1}{L} \sum_{\ell=1}^{L} CN_i(B_i^{(\ell)}), \tag{4}$$

where $B_i^{(\ell)}$ denotes the ℓ-th batch containing sample i. The stratification preserves the composition of the evaluation dataset across batches, while repeated reshuffling exposes each sample to different sets of competing candidates. We assess the convergence of this estimator with the number of repetitions in Section A.3.

3.2 Model-level Modality Gap

To quantify the modality gap at the model level, we use the *maximum robust offset score* (MROS), which measures how strongly the largest modality-specific mean offsets stand out from the remaining representation. Unlike aggregate separation measures such as L2M (Liang et al., 2022) and RMG (Schrodi et al., 2025), which can already attain substantial values at initialization, MROS specifically captures the sparse outlier structure characteristic of the trained modality gap. The full definition is provided in Section A.2.

4 Understanding the Modality Gap

Experimental setup. We evaluate models on a fixed subset of approximately 20M image–text pairs from a 10% subset of LAION-400M (Schuhmann et al., 2021) for which the original image URLs remained accessible, referred to as LAION-20M*. We additionally evaluate on the COCO validation set (Lin et al., 2014). We study OpenCLIP models from the scaling-law suite (Cherti et al., 2023) with intermediate checkpoints, together with a broader set of publicly available pretrained models.

4.1 A Mechanistic Perspective on the Modality Gap

We begin by examining how the modality gap emerges during training. Figure 2 (left) traces the dimension-wise image and text means of a representative CLIP model across training (i.e., model 80M/34B/88k; Table 1 for details; see Figure 10 and Figure 11 for different models with intermediate checkpoints). At the beginning of optimization, the means of both modalities are concentrated around the origin and no individual dimension exhibits a pronounced modality-specific offset. Within the first training epochs, however, a small number of dimensions separate sharply from the remaining representation: some develop a large image-side offset mean while their text mean remains close to zero, whereas others exhibit the converse pattern. These offsets strengthen during early and intermediate training and remain identifiable at convergence, although prolonged training can reduce their absolute magnitude. Across the pre-trained models we study, the exact number and strength of these dimensions varies, but at least one pronounced offset consistently appears on both the image and text side (see Figure 12 for different pre-trained models, see Figure 13 for different backbones). Notably, the smallest and least-trained models can exhibit little or no such outlier structure, suggesting that the characteristic modality gap emerges only after sufficient contrastive training on sufficiently large datasets. Models trained on larger datasets also tend to reach higher MROS, consistent with greater expected exposure to correspondence noise during training.

Having established how the modality gap emerges, we next examine the functional role of the dimensions driving it. Similarity in CLIP is the sum of dimension-wise cross-modal products, such that dimension g contributes $x_{ig}y_{jg}$ to the similarity between image i and text j. Consider an image-side gap dimension (i.e., a large image-side offset). For a fixed image query i, its GT retrieval rank depends on whether the paired text i is more similar than each competing text j within the retrieval batch. The contribution of dimension g to this GT–competitor similarity margin is $x_{ig}(y_{ig} - y_{jg})$. Because the offset places x_{ig} systematically away from zero, differences between the paired and competing texts along dimension g are scaled by a comparatively large factor. Gap dimensions can therefore exert disproportionate influence on whether the GT match outranks individual competitors and, consequently, on its retrieval rank. Importantly, this effect is determined by the interaction of both modalities rather than by either coordinate in isolation. Figure 1 illustrates this mechanism for a query whose paired caption overtakes a competing caption once the gap contribution is included.

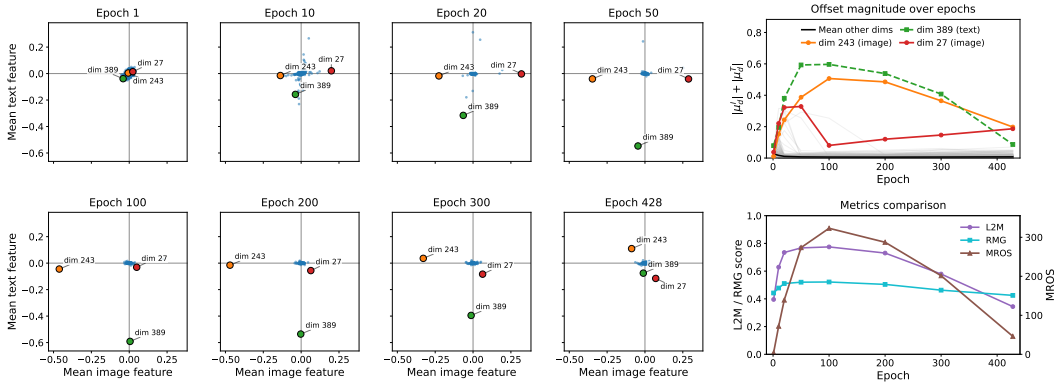

Figure 2: Emergence of the modality gap during training for model 80M/34B/88k. Left: means of image and text representations across training epochs. Right top: evolution of per-dimension cross-modal sum of absolute offsets. Right bottom: comparison of different modality gap metrics.

We illustrate this behavior on a fixed shard of 20,000 paired LAION-20M* samples for model 80M/34B/88k at convergence, keeping the retrieval candidate set constant across interventions. For image-to-text retrieval, each image serves as a query, while the texts are ranked by cosine similarity, with the paired text defining the GT; text-to-image retrieval is evaluated analogously. For each dimension, we remove its coordinate from the representations and renormalize the remaining coordinates to unit length before recomputing similarities, preserving cosine normalization after dimensional removal. This choice has little influence on the resulting sample rankings: omitting renormalization yields highly similar results ($\rho = 0.96$); Section A.4). We then record the absolute change in GT rank relative to the unmodified representation. Modality-gap dimensions stand out sharply under this intervention: while most dimensions induce only very small GT rank changes, removing a gap dimension produces substantially larger changes (Figure 3, top). These effects are nevertheless highly concentrated on a few samples. Even under a threshold of $|\Delta\text{GT-rank}| > 1$, at most about 20% of samples are affected in their GT rank by any gap dimension across training, falling below 10% toward convergence (Figure 3, bottom). The few non-gap dimensions with larger GT-rank effects primarily act alongside gap dimensions on the same samples: after excluding pairs with $|\Delta\text{GT-rank}| \geq 3$ under any gap-dimension ablation, their larger GT-rank effects disappear (Table 2). This concentration also makes the analysis relatively insensitive to the exact set of gap dimensions selected.

The sample-specific re-ranking effect is determined by the cross-modal product rather than by the magnitude of the offset-side coordinate alone. Across three gap dimensions, correlations between coordinate magnitude and absolute GT rank change remain weak (Spearman $|\rho| = 0.06$–0.24; Figure 14). Further, the gap dimensions affect the same samples: among the top 2% most affected pairs, pairwise overlap ranges from 79–89% for image-to-text and 81–91% for text-to-image retrieval (Table 3). Together, these results indicate that modality-gap dimensions jointly drive a highly sample-specific re-ranking mechanism.

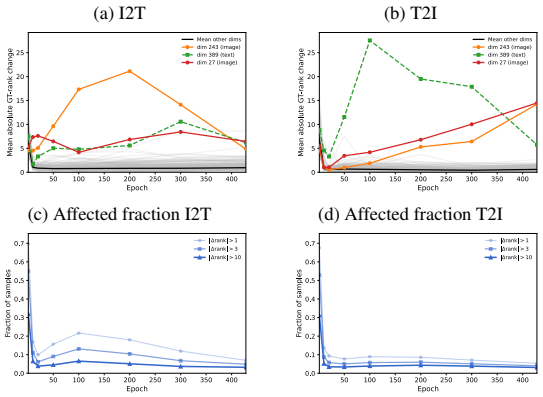

Figure 3: Functional effect of modality-gap dimensions across training. Top: mean absolute GT rank change after dimensional ablation. Bottom: fraction of samples exceeding different absolute GT rank-change thresholds in at least one modality-gap dimension.

To quantify this sample-specific sensitivity at scale, we use the *CN score* introduced in Section 3.1, which aggregates the absolute GT-rank changes induced by gap-dimension ablation across dimensions and retrieval directions. We compute the score for the full LAION-20M* dataset using batches of $M = 5{,}000$ samples and $L = 200$ repetitions, such that

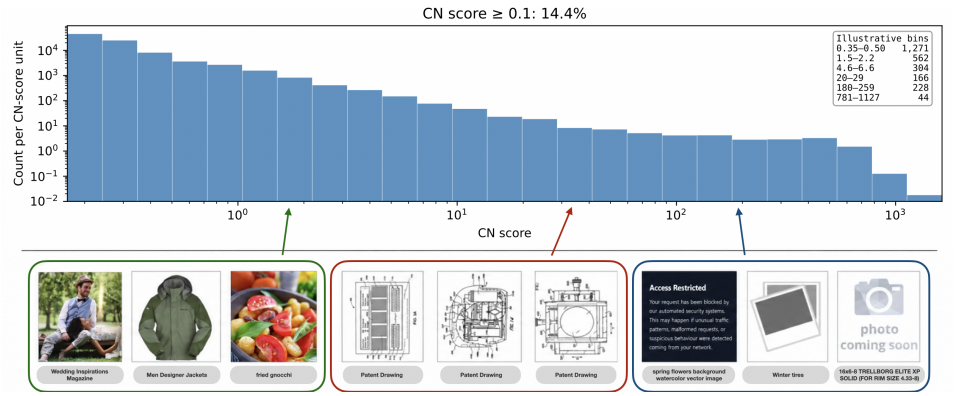

Figure 4: Distribution of *CN scores* for model 80M/34B/88k. Green highlights weak or incomplete correspondence, red underspecified one-to-many pairs, blue severe correspondence noise.

each sample is evaluated in 200 different retrieval contexts (see Section A.3 for robustness check). The resulting distribution is strongly long-tailed: approximately 85% of samples have $CN < 0.1$, indicating only negligible average GT-rank changes across modality-gap dimensions. We visualize the remaining tail in Figure 4. A qualitative examination of samples along this tail shows that pairs with lower *CN scores* increasingly exhibit weak or incomplete image–text correspondence (in green), whereas larger *CN scores* are dominated by severe correspondence failures, such as missing or malformed captions and images with image-not-available placeholder (in blue), as well as strongly underspecified one-to-many correspondences such as the "Patent Drawing" example (in red). In contrast, inspected samples with $CN < 0.1$ exhibit close image–text correspondence. Additional qualitative examples are shown in Figure 15.

4.2 Causal Effects of Correspondence Noise

The previous section showed that higher *CN scores* are associated with weaker or erroneous image–text correspondence in existing LAION data. We test this relationship causally by intervening on fixed image–text pairs using four forms of correspondence noise: (1) text corruption, (2) image corruption, (3) textual information ablation, and (4) visual information ablation. Types 1/2 apply predominantly parameterized corruptions to ten COCO image–caption pairs, whereas Types 3/4 use ten pre-selected, visually rich internet images and induce correspondence noise through modality-specific information ablation. The latter involve subjective, scene-specific choices; full corruption functions, parameterizations, and ablations are provided in the supplemental material (Section D.1)

Types 1–2: Text and image corruption. We corrupt either the caption or image while holding the paired modality fixed. Text corruptions span eight families: out-of-distribution one-word replacement, in-corpus five-word replacement, character noise, word-order shuffling, generic *image unavailable* captions, hallucinated extensions, repetition artifacts, and static or caption-conditioned web/parsing artifacts. Image corruptions likewise span eight families: patch shuffling, bilateral filtering, Gaussian blur, grayscale conversion, channel shuffling, CutOut, Gaussian noise, and "image not available" placeholders. We generate multiple variants per corruption, varying strength or realization where applicable.

Types 3–4: Modality-specific information ablation. For each visually rich image, we construct a detailed reference caption and ten progressively information-ablated but still correct captions, varying how much visual information is retained in text without introducing an incorrect caption. Conversely, holding the detailed caption fixed, we select five semantically meaningful image regions and remove each partially or fully, yielding ten targeted visual ablations per image.

We evaluate all manipulations in two retrieval environments. In COCO, each manipulated focal pair is embedded among the remaining 4,999 validation pairs. For LAION-20M*, we sample 499,900 unique pairs and partition them into 100 disjoint contexts of 4,999 negatives each. Clean and manipulated versions of each focal pair are evaluated against the same 100 batches of 4,999 negatives, yielding 100 retrieval evaluations per condition. Our primary outcome is the change in *CN score*, $\Delta CN_i = CN_{i,\text{corrupt}} - CN_{i,\text{clean}}$.

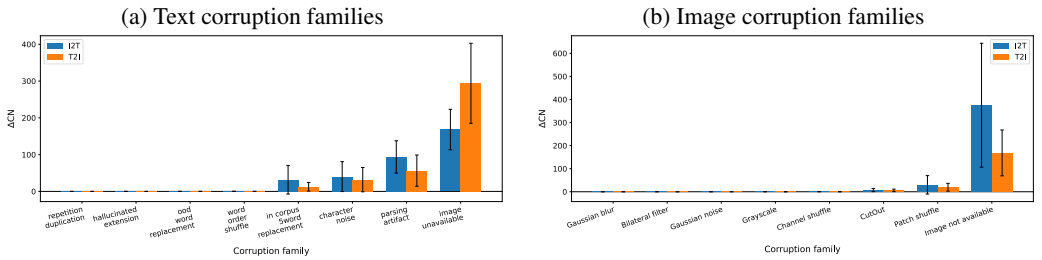

Figure 5: Effects of controlled text and image corruptions on correspondence sensitivity across text- and image-corruption families within COCO.



Original caption: A woman in a blue dress prepares food on a kitchen island beside bread, cheese, and a jar, while an orange cat sits on the counter near a kettle, with hanging pans, cabinets, dishes, and a double sink behind them.

Figure 6: Controlled reduction of correspondence information (Type 3/4) for LAION.

Figure 5 summarizes interventions for Types 1/2 in COCO (full and LAION-20M* results in Figure 17 and Figure 18). Text corruptions produce the largest increases in *CN score* when captions become largely uninformative about the image: generic *image unavailable* captions and parsing artifacts without image-specific content show the strongest effects. Partially informative corruptions are weaker: keyword-preserving parsing artifacts remain near the clean baseline, character noise increases ΔCN mainly at high severity, and perturbations preserving substantial caption information (e.g., repetitions or OOD word replacement) have little effect. Image corruptions show similar tendencies: the image-not-available placeholder produces the strongest increase in ΔCN, patch shuffling and CutOut produce clear but more moderate effects, whereas texture or color removal induce little change. Stronger interventions generally yield larger changes, indicating that sensitivity to modality-gap dimensions rises most when correspondence is severely degraded or absent.

Figure 6 illustrates Types 3/4 for one visually rich image in LAION-20M* (additional results in Figure 19 and Figure 20). Reducing textual information generally produces substantially larger ΔCN than visual ablation: increasingly general captions increase *CN scores*, consistent with prior findings that reducing caption information enlarges the modality gap (Schrodi et al., 2025). Most targeted visual removals have little effect unless large or central regions are removed, such as the woman in the kitchen. Together, these interventions indicate that the influence of modality-gap dimensions increases as cross-modal correspondence becomes less specific.

4.3 Consistency of Correspondence Noise Rankings

We next test whether *CN score* rankings are stable over training and across independently trained CLIP and SigLIP models. For every checkpoint and model, we compute *CN scores* on the same fixed set of 100,000 LAION image–text pairs using 40 repeated retrieval contexts of 5,000 samples each, and compare the resulting rankings using Spearman's rank correlation.

Intra-model consistency. We first follow the OpenCLIP model 80M/34B/88k across 11 checkpoints from epoch 1 to 428 (see Figure 7, left). Early in training, *CN score* rankings change substantially, with weak agreement between the first checkpoints and later stages. However, as the modality

gap emerges and reaches its most pronounced levels around epoch 100 (see Figure 2), the ranking becomes markedly more stable. This suggests that once the modality gap is established, the relative ordering of samples by *CN score* is comparatively persistent throughout the remainder of training.

Inter-model consistency. Next, we compare the same sample rankings across ten CLIP models and one SigLIP model. Despite substantial differences in training data, optimization, and the structure of the modality-gap dimensions, most model pairs exhibit moderate positive rank correlations, only slightly lower than those observed across checkpoints of the same model. This suggests that the *CN score* ranking captures a shared property of image–text pairs that is not specific to a particular training run or dataset.

Two models deviate from this pattern. DFN (Fang et al., 2024) shows weaker agreement with the remaining models, while EntityNet (Ging et al., 2025) is almost uncorrelated with them. Both models differ substantially in how their training data are curated: DFN is trained on 2B pairs selected by a learned filtering network from a 12.8B-pair CommonPool, whereas EntityNet constructs its dataset through targeted, knowledge-graph-guided web searches and entity-based text supervision rather than broad web-scale pair collection. These curation strategies plausibly reduce severe correspondence noise. Notably, EntityNet nevertheless exhibits a modality gap (Figure 12), indicating that such noise is not necessary for the gap to emerge. Consistently, *CN score* rankings computed on EntityNet training samples generalize across models (Figure 16), suggesting that a shared component reflects milder correspondence mismatches, while stronger agreement on broadly scraped web data partly reflects more severe noise patterns.

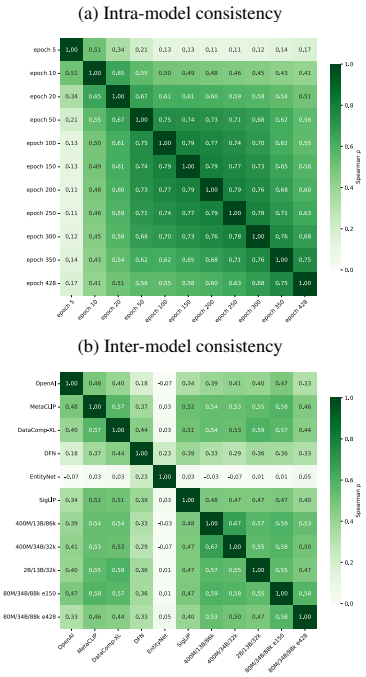

Figure 7: Consistency of *CN score* rankings. Spearman correlations across **(a)** checkpoints model 80M/34B/88k; **(b)** independently trained CLIP/SigLIP models.

4.4 Filtering Correspondence Noise Reduces the Modality Gap

To test whether correspondence noise contributes to the emergence of the modality gap during training, we train two ViT-B/32 CLIP models on 10M image–text pairs. The low-CN model uses pairs with $CN < 0.01$, whereas the mixed-CN model uses an equal mixture of pairs with $CN < 0.01$ and $0.01 \leq CN < 30$. Filtering scores were estimated from only five retrieval batches, as these training runs were initiated before the more stable CN estimates used in our subsequent analyses were available. Pairs with $CN \geq 30$ are excluded from both datasets to limit extreme correspondence failures. Both models are trained for 200 epochs under otherwise identical conditions (Section A.5). We track MROS and evaluate ImageNet 0-shot accuracy and COCO I2T/T2I R@1.

Filtering correspondence noise reduces both peak and converged modality gap (Figure 8): the mixed-CN model reaches higher peak and final MROS, although the low-CN model is slightly higher at two intermediate checkpoints. The low-CN model also achieves stronger downstream performance. A substantial modality gap nevertheless remains under low-CN filtering. Because the filtering scores were estimated from only five retrieval contexts, this residual gap may partly reflect remaining correspondence noise (Section A.6). Because the intervention acts on correspondence noise rather than the gap itself, we interpret the reduced gap and improved performance as parallel consequences of cleaner correspondence, rather than evidence that reducing the gap directly improves performance.

4.5 Modality Gap and Downstream Performance

To assess whether modality-gap magnitude is informative about representation quality, we compare MROS with downstream performance across training checkpoints and 21 fully trained models. Across training checkpoints, MROS shows moderate positive correlations with ImageNet 0-shot

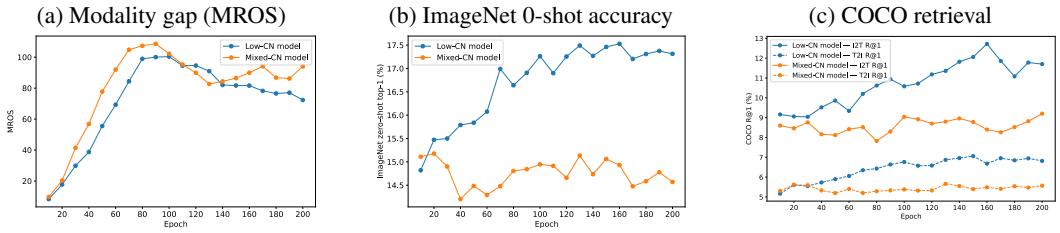

Figure 8: Effect of correspondence noise filtering before training. Left: evolution of the modality gap measured by MROS. Middle: ImageNet 0-shot performance. Right: COCO I2T and T2I R@1.

accuracy and COCO retrieval ($r = 0.524$–0.555; Figure 22). Across converged models, the association remains moderate for ImageNet accuracy ($r = 0.564$, $p = 0.008$), but weaker for COCO I2T and T2I Recall@1 ($r = 0.317$, $p = 0.161$; $r = 0.328$, $p = 0.147$; Figure 23). This does not necessarily contradict the filtering experiment: there, correspondence noise is varied under an otherwise matched training setup, whereas cross-model comparisons additionally reflect differences in training data and sample exposure. Greater training scale can simultaneously improve downstream performance and increase exposure to imperfect image–text correspondences, which may in turn produce a more pronounced modality gap despite improving representation quality. Thus, gap magnitude is not a simple proxy for representation quality; its relationship to performance depends on the factors producing it (e.g., amount of weak/noisy correspondence, size of clean training data).

We next revisit the functional importance of gap dimensions. Schrodi et al. (2025) showed that zeroing the modality-gap dimensions can sharply reduce downstream performance, attributing this effect to the loss of rich semantic information. We replicate and extend this analysis for model 80M/34B/88k at epochs 214 and 428 and model 400M/13B/86k at epochs 16 and 32, ablating either the top two or top four gap dimensions (Figure 24 and Figure 25). We then replace the offset-side values of all samples by the modality mean, removing sample-specific information (i.e., variation) while preserving the systematic offset. Across these settings, Schrodi-style ablation produces the strongest performance losses when the modality gap is most pronounced, particularly for ImageNet and COCO I2T, whereas mean replacement recovers most of the lost performance and generally approaches the random-dimension control. By contrast, renormalizing after either intervention again induces substantially larger losses. Together, these results suggest that the disproportionate ablation effect of gap dimensions primarily reflects geometric calibration around their large modality-specific offsets, rather than unusually rich sample-specific semantic information.

5 Conclusion

We studied the modality gap from a mechanistic perspective and showed that it is closely tied to imperfect image–text correspondence. The gap is concentrated in a small number of high-offset dimensions whose influence on similarity rankings is highly sample-specific and strongest for pairs with weak or noisy correspondence. We quantify this sample-specific dependence with the *CN score*, while controlled text and image interventions provide causal evidence that the influence of gap dimensions increases as correspondence degrades. *CN score* rankings remain substantially consistent across checkpoints and models, and filtering correspondence noise before training jointly reduces the gap and improves downstream performance. Yet, the gap magnitude is positively but inconsistently associated with performance across converged models, showing that it is not a simple proxy for representation quality; gap-dimension ablations further suggest that their apparent importance largely reflects the geometry induced by their offsets. Overall, the modality gap appears to emerge, at least in part, as contrastive models adapt to noisy image–text correspondence, rather than as a consistent structural misalignment that should simply be minimized.

Limitations. Our results do not imply that correspondence noise fully explains the modality gap: even training exclusively on low *CN score* data reduces rather than eliminates it. Moreover, our controlled training experiment uses only two runs on a comparatively small 10M-sample subset, motivating larger-scale studies that disentangle correspondence quality from factors such as the number of clean training pairs and the prevalence and severity of weak or erroneous correspondences.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the implementation of research code used for analysis. We did not use generative AI tools to generate synthetic datasets, develop theoretical or conceptual frameworks, formulate or prove mathematical claims, propose hypotheses, design the research methodology or experiments, perform qualitative analysis, or interpret the empirical results; the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to identify potentially relevant literature missed in our manual literature search, polish the writing for clarity and concision, provide critical feedback on an initial manuscript draft, and assist with the design and generation of the central schematic figure.

All AI-assisted work was reviewed by the authors. AI-assisted code was inspected and tested against the intended analysis, suggested literature was manually verified before inclusion, and all manuscript edits, reviewer-style feedback, and figure content were critically evaluated and revised as appropriate. The authors take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies existing image–text datasets and publicly available pretrained models and does not involve interaction with human participants. Some experiments use web-sourced image–text data and publicly accessible images, which may inherit privacy, copyright, and representational issues present in web-scale datasets. The proposed *CN score* measures model-dependent image–text correspondence rather than intrinsic sample quality; its use for dataset filtering should therefore consider potential changes in dataset composition and associated biases.

REPRODUCIBILITY STATEMENT

We provide the definitions and procedures required to reproduce our analyses in Sections 3–4 and supplementary methodological details in Appendix A, including the exact pretrained model identifiers (Appendix A.1), modality-gap metric (Appendix A.2), convergence analysis of the CN score estimator (Appendix A.3), and training configuration for our correspondence-noise filtering experiment (Appendix A.5). The corruption functions, parameterizations, and modality-specific ablations used in the causal experiments are documented in Appendix D.1. Most experiments use publicly available pretrained models, including OpenCLIP models with explicitly reported checkpoint identifiers; the two models trained for this work follow the standard OpenCLIP training pipeline under identical training settings apart from dataset composition. Upon acceptance, we will release the code required to compute the proposed metrics, reproduce the analyses and figures, and run the corresponding training experiments. Experiments based on COCO can be reproduced using the publicly available dataset. Exact reconstruction of our LAION-20M* subset is inherently limited by the dynamic nature of LAION-400M, which references images through external web URLs: approximately 40% of retrieval requests already failed when constructing our dataset, and URL availability can change over time. Consequently, the exact image set may no longer be recoverable, although our released code will reproduce the sampling and analysis procedure on the subset of LAION data that remains accessible.

REFERENCES

Na Min An, Eunki Kim, James Thorne, and Hyunjung Shim. I0T: Embedding Standardization Method Towards Zero Modality Gap. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 27182–27199, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1319. URL https://aclanthology.org/2025.acl-long.1319/.

Tom Burgert, Oliver Stoll, Paolo Rota, and Begüm Demir. ImageNet-trained CNNs are not biased towards texture: Revisiting feature reliance through controlled suppression. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829. IEEE, 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*, pp. 30288–30298, June 2026.

Zhuohang Dang, Minnan Luo, Chengyou Jia, Guang Dai, Xiaojun Chang, and Jingdong Wang. Noisy correspondence learning with self-reinforcing errors mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1463–1471, 2024.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in CLIP. In *International Conference on Learning Representations*, volume 2025, pp. 82070–82088, 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.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *International Conference on Learning Representations*, volume 2024, pp. 36221–36237, 2024.

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 Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei W Koh, Olga Saukh, Alexander J 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. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Simon Ging, Sebastian Walter, Jelena Bratulić, Johannes Dienert, Hannah Bast, and Thomas Brox. Using Knowledge Graphs to harvest datasets for efficient CLIP model training. In *DAGM German Conference on Pattern Recognition*, pp. 287–302. Springer, 2025.

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. URL https://openreview.net/forum?id=RHPqr2eqJO.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. Mind the Gap: Preserving and Compensating for the Modality Gap in CLIP-Based Continual Learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 3777–3786, 2025. doi: 10.1109/ICCV51701.2025.00360.

Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021.

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 *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 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 Multi-Modal Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7661–7671, June 2023.

Binxu Li, Yuhui Zhang, Xiaohan Wang, Weixin Liang, Ludwig Schmidt, and Serena Yeung-Levy. Closing the Modality Gap for Mixed Modality Search. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=jKylyDMFcs.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Fabian Mager, Hiba Nassar, and Lars Kai Hansen. On the modality gap and the contrastive loss in multi-modal representation learning. *arXiv preprint arXiv:2607.10698*, 2026.

Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew Bagdanov. Cross the gap: Exposing the intra-modal misalignment in clip via modality inversion. In *International Conference on Learning Representations*, volume 2025, pp. 77750–77771, 2025.

Changdae Oh, Junhyuk So, Hoyoon Byun, YongTaek Lim, Minchul Shin, Jong-June Jeon, and Kyungwoo Song. Geodesic multi-modal mixup for robust fine-tuning. *Advances in Neural Information Processing Systems*, 36:52326–52341, 2023.

Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal active complementary learning with self-refining correspondence. *Advances in Neural Information Processing Systems*, 36:24829–24840, 2023.

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 *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the Modality Gap: An Exploration in the Hyperbolic Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27263–27272, June 2024.

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 *International Conference on Learning Representations*, volume 2025, pp. 27836–27864, 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. In *NeurIPS Workshop on Data-Centric AI*, 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.

Wuyuqing Wang, Zeyuan Gu, and Erkun Yang. Noisy correspondence learning with modality gap direction correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10163–10171, 2026a.

Yimu Wang, Evelien Riddell, Adrian Chow, Sean Sedwards, and Krzysztof Czarnecki. Mitigating the Modality Gap: Few-Shot Out-of-Distribution Detection with Multi-modal Prototypes and Image Bias Estimation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2741–2751, 2026b. doi: 10.1109/WACV61042.2026.00269.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *International Conference on Learning Representations*, volume 2024, pp. 47812–47831, 2024.

Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. Post-pre-training for Modality Alignment in Vision-Language Foundation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4256–4266, June 2025.

Mouxing Yang, Yunfan Li, Zhenyu Huang, Zitao Liu, Peng Hu, and Xi Peng. Partially View-Aligned Representation Learning With Noise-Robust Contrastive Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1134–1143, June 2021.

Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. BiCro: Noisy Correspondence Rectification for Multi-Modality Data via Bi-Directional Cross-Modal Similarity Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19883–19892, June 2023.

Yuchen Yang, Likai Wang, Erkun Yang, and Cheng Deng. Robust Noisy Correspondence Learning with Equivariant Similarity Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17700–17709, June 2024.

Can Yaras, Siyi Chen, Peng Wang, and Qing Qu. Explaining and Mitigating the Modality Gap in Contrastive Multimodal Learning. In Beidi Chen, Shijia Liu, Mert Pilanci, Weijie Su, Jeremias Sulam, Yuxiang Wang, and Zhihui Zhu (eds.), *Conference on Parsimony and Learning*, volume 280 of *Proceedings of Machine Learning Research*, pp. 1365–1387. PMLR, March 2025. URL https://proceedings.mlr.press/v280/yaras25a.html.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, October 2023.

Yuhui Zhang, Jeff Z. HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and Rectifying Vision Models using Language. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=D-zfUK7BR6c.

Zihua Zhao, Mengxi Chen, Tianjie Dai, Jiangchao Yao, Bo Han, Ya Zhang, and Yanfeng Wang. Mitigating Noisy Correspondence by Geometrical Structure Consistency Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27381–27390, June 2024.