UNDERSTANDING AND CORRECTING CONFORMITY IN INTRA-MODAL RETRIEVAL

Anonymous authors
Paper under double-blind review

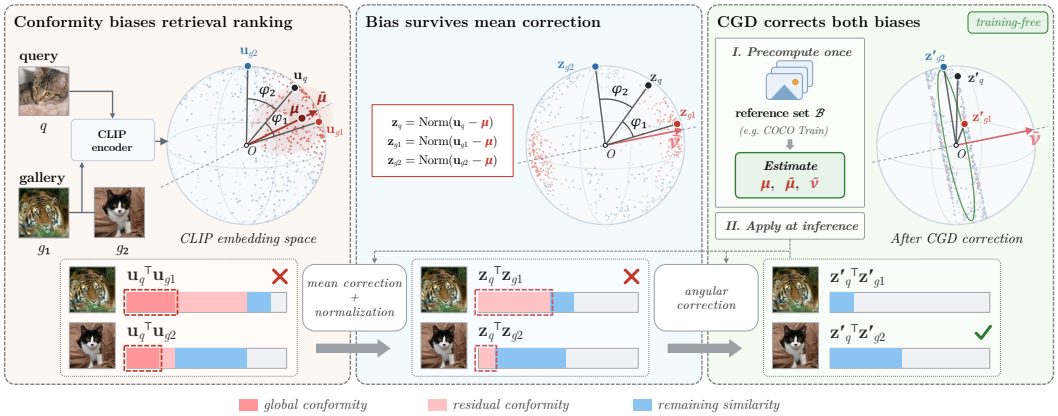


Figure 1: **Overview of Conformity-Guided Debiasing (CGD). Left:** Conformity is captured by a shared direction $\bar{\boldsymbol{\mu}}$ whose contribution to intra-modal cosine similarity can bias retrieval rankings. **Middle:** Subtracting the reference mean $\boldsymbol{\mu}$ and re-normalizing the embeddings, $\mathbf{z}_i = \mathrm{Norm}(\mathbf{u}_i - \boldsymbol{\mu})$, removes global conformity, but residual conformity remains after projection back onto the unit sphere. **Right:** CGD estimates $\boldsymbol{\mu}$, the conformity direction $\bar{\mathbf{u}}$, and the residual conformity direction $\bar{\boldsymbol{\nu}}$ once from an external reference bank $\mathcal{B}$. At inference, it applies the fixed correction to each embedding: after mean correction, the signed residual projection $\beta_i = \mathbf{z}_i^\top \bar{\boldsymbol{\nu}}$ defines the angle $\theta_i = \arcsin \beta_i$ from the subspace orthogonal to $\bar{\boldsymbol{\nu}}$, which is contracted to $\theta_i' = t_i \theta_i$. Removing *both global and residual conformity* restores the correct retrieval ranking, while requiring no training or query/gallery fitting.

ABSTRACT

Frozen vision–language representations are widely used for intra-modal retrieval, yet their embedding geometry contains systematic shared structures that can distort cosine similarity and degrade ranking quality. We first show that conformity, the tendency of an embedding to be similar to many samples, is exactly determined by the embedding's projection onto a single shared direction defined by the normalized reference mean. We further show that mean correction alone is insufficient: after subtracting the reference mean, the L_2-normalization implicitly imposed by cosine similarity in retrieval gives rise to a distinct residual conformity direction. Motivated by these observations, we propose Conformity-Guided Debiasing (CGD), a training-free geometric correction that attenuates conformity and residual conformity. CGD first applies mean correction and L_2-normalization, and then performs a sample-dependent angular contraction toward the subspace orthogonal to the residual conformity direction. The resulting transform is plug-and-play, requires no backbone modification or query gallery fitting, and can be applied on top of existing intra-modal retrieval methods. Across multiple CLIP-like backbones and image-to-image and text-to-text benchmarks, our method consistently improves retrieval performance, showing that explicitly correcting conformity-related directional structure provides a simple and effective way to refine frozen intra-modal representations.

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.

1 INTRODUCTION

Vision-Language Models (VLMs), such as CLIP (Radford et al., 2021), have become widely used as general-purpose representations for large-scale multimedia retrieval. These models map images and text into a shared high-dimensional representation space that can support multiple retrieval tasks, including cross-modal retrieval and intra-modal retrieval, without maintaining separate task-specific embedding spaces. Since CLIP is trained to optimize vision-language alignment rather than intra-modal similarity, specialized unimodal representations may provide stronger performance for image-to-image (I2I) and text-to-text (T2T) retrieval, but the shared representations are particularly useful for I2I and T2T retrieval when the same representations are expected to support multiple contemporaneous retrieval modes Zhu et al. (2024); Liang et al. (2025); Beal et al. (2026).

Understanding the properties of VLM-based similarity spaces is therefore of significant interest for large-scale multimedia retrieval. Recent studies of the geometry of CLIP representations Levi & Gilboa (2025); Liang et al. (2022); Lin et al. (2025) reveal that some samples tend to have relatively high similarity to many samples at the same time. The extent to which a sample exhibits this behavior can be quantified by its average cosine similarity to samples from a representative dataset. Levi & Gilboa (2025) refer to this quantity as *conformity* and show that it can be accurately estimated from the sample's alignment with the modality mean. Therefore, a sample with low conformity is close to only a few semantically related neighbors, while a sample with high conformity tends to have relatively high similarity to many samples at the same time. If this shared similarity does not correspond to an actual semantic relation, it may act as a common similarity component that is unrelated to query-specific semantics, influencing many query-gallery comparisons and disturbing the semantic signals that should otherwise determine the final ranking. Our work therefore explores the following hypothesis: *conformity in CLIP representations systematically and negatively affects intra-modal retrieval quality.*

To evaluate the impact of conformity, we must first turn conformity from a descriptive population-level statistic into a geometric object that can be precisely analyzed and manipulated. We show that this can be done exactly: for any fixed L_2-normalized reference set, the notion of conformity can be fully reduced to the projection of queries onto a single shared direction, which we term as *conformity direction*, given by the normalized modality mean estimated from that reference set. The reference set is a collection of same-modality samples that are disjoint from both the query and gallery sets. When query embeddings are close to this direction, conformity contributes significantly to their cosine similarity to many gallery embeddings, artificially boosting the scores of those embeddings.

As shown in Fig. 1, the connection between conformity and the modality mean suggests mean centering as an immediate correction. Although centering removes the original mean component, cosine-based retrieval L_2-normalizes the centered representations, after which they exhibit a new shared direction distinct from the original conformity direction. We define this remaining structure, which can still boost the ranking of certain gallery images, as *residual conformity*.

These observations motivate *Conformity-Guided Debiasing (CGD)*, which progressively removes both primary and residual conformity by remapping input representations into a more discriminative space, guided by a pre-corrected reference bank. CGD is a training-free method that first corrects for the conformity direction, then re-estimates the residual conformity axis in the new normalized representation and adaptively applies bidirectional angular contraction according to the absolute projection strength of each sample along this axis. We show that CGD only requires a small reference set for estimating the conformity direction, requires no backbone training and does not use target labels or query/gallery statistics, and it can be applied directly and efficiently to the frozen representations as a plug-and-play correction. Our contributions are summarized as follows:

- We reduce conformity exactly to a shared direction and show that it directly impacts the intra-modal retrieval score.
- Our study reveals that retrieval-relevant residual conformity remains after global correction, with a clear bidirectional structure.
- We propose CGD, a training-free, reference-only progressive correction process for both global and residual conformity.
- CGD operates in a plug-and-play manner, achieving consistent retrieval gains on top of state-of-the-art methods.

2 Related Work

2.1 Geometry of Contrastive Vision-Language Representations

Dual-encoder vision-language models such as CLIP (Radford et al., 2021) map images and texts into a shared embedding space through contrastive learning, where cosine similarity is used as the basic similarity measure. Subsequent studies have shown that this shared space exhibits clear non-uniform geometric structures. Prior work has analyzed the image-text modality gap (Liang et al., 2022), information imbalance across modalities (Schrodi et al., 2025), the ellipsoidal geometry of CLIP representations (Levi & Gilboa, 2025), and text-based decomposition of visual features (Gandelsman et al., 2024). IsoCLIP (Magistri et al., 2026) further studies the joint operator formed by the visual and textual projectors and reveals a strongly non-uniform singular spectrum, with a relatively flat middle region and more anisotropic directions at both spectral extremes. These studies establish that CLIP representations contain substantial non-uniform geometric structure, but largely characterize such structure through global distributional properties or population-level statistics. Our work instead studies its direct role in intra-modal similarity. Specifically, we show that, on the unit hypersphere, conformity reduces exactly to projection onto a single shared direction, allowing its contribution to pairwise cosine similarity and retrieval rankings to be isolated and functionally analyzed.

2.2 Intra-Modal Similarity Correction

Frozen CLIP representations are also widely used for image-to-image and text-to-text retrieval, yet different works provide different explanations. Cross-the-Gap (Mistretta et al., 2025) attributes the suboptimality of raw intra-modal cosine similarity to intra-modal misalignment and uses modality inversion to convert intra-modal comparison into cross-modal comparison. IsoCLIP (Magistri et al., 2026) follows this view and derives an inter-modal operator from the frozen projectors, retaining its middle spectral region while suppressing directions at both spectral extremes. In contrast, Herzog & Wang (2026) revisit the intra-modal misalignment hypothesis, arguing that image-image similarities do not have the previously assumed arbitrary degrees of freedom. They propose PCA$\leftarrow$, which projects image representations onto a leading PCA subspace derived from ImageNet class-name text embeddings. Together, prior methods improve intra-modal similarity through modality inversion, spectral filtering, or subspace projection. We directly model conformity as a source of shared similarity and apply a simple post-hoc correction estimated from an external reference set. Our method can be applied to any frozen CLIP-like model alone or combined with existing corrections such as IsoCLIP and PCA$\leftarrow$.

3 Revisiting Conformity in Intra-Modal Retrieval

3.1 Background on Conformity

Levi & Gilboa (2025) show that the embeddings of each CLIP modality concentrate near a tilted ellipsoidal shell displaced from the origin and to quantify this phenomenon, they define the *conformity* of an embedding as its expected cosine similarity to the other same-modality samples in a *reference set* S that is disjoint from both the query and gallery sets. For $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ of N samples, the conformity of $\mathbf{v}_j \in S$ is defined as:

$$C(\mathbf{v}_j) = \mathbb{E}_{\mathbf{v}_k \in S,\, k \neq j} \left[\cos(\mathbf{v}_j, \mathbf{v}_k)\right]. \tag{1}$$

The direct evaluation of $C(\mathbf{v}_j)$ requires averaging pairwise similarities over S. It can instead be approximated via a surrogate based on the similarity between an embedding and the modality mean, $\widehat{C}(\mathbf{v}_j) = a\cos(\mathbf{m}, \mathbf{v}_j) + b$ with $\mathbf{m} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{v}_i$, with constants a, b fitted per encoder and dataset.

Thus, conformity of a generic element *outside* the reference set is captured mostly by an embeddings alignment with the modality mean of the reference set, distinguishing generic content (high conformity) from rare, distinctive concepts (low conformity). This relationship is derived in the unnormalized embedding space, whereas retrieval operates on the unit hypersphere $\mathbb{S}^{d-1}$ through cosine similarity. We next show that on $\mathbb{S}^{d-1}$ the mean-based surrogate becomes an exact identity, with conformity governed entirely by the projection onto a single direction.

3.2 CONFORMITY ON THE UNIT HYPERSPHERE

Let $\mathbf{x} \in \mathbb{R}^d$ denote the representation produced by a frozen encoder and let $\mathbf{u} = \mathrm{Norm}(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|}$ be the projection of $\mathbf{x}$ onto the unit-sphere.

Lemma 1 (Conformity on the unit hypersphere)**.** *Let the reference set $\mathcal{B} = \{\mathbf{u}_i\}_{i=1}^M \subset \mathbb{S}^{d-1}$, $M \geq 2$, have mean $\boldsymbol{\mu} = \frac{1}{M}\sum_k \mathbf{u}_k \neq \mathbf{0}$ and direction $\overline{\boldsymbol{\mu}} = \mathrm{Norm}(\boldsymbol{\mu})$. Since $\sum_{k\neq i} \mathbf{u}_k = M\boldsymbol{\mu} - \mathbf{u}_i$ and $\mathbf{u}_i^\top \mathbf{u}_i = 1$, the conformity Eq. (1) of $\mathbf{u}_i$ in $\mathcal{B}$ is*

$$C(\mathbf{u}_i) = \frac{\mathbf{u}_i^\top (M\boldsymbol{\mu} - \mathbf{u}_i)}{M-1} = \frac{M\|\boldsymbol{\mu}\|}{M-1}\,\alpha_i - \frac{1}{M-1}, \qquad \alpha_i = \mathbf{u}_i^\top \overline{\boldsymbol{\mu}}, \tag{2}$$

a strictly increasing affine function of α_i.

We refer to $\overline{\boldsymbol{\mu}}$ as the ***conformity direction***. As shown in Fig. 2 (right-(i)), samples with larger projection α_i have higher conformity. Thus, $\overline{\boldsymbol{\mu}}$ captures a global commonality direction in the encoder's representation space, with generic content tending to project stronger onto it and rarer content weaker. In practice, this direction is estimated from a reference set of samples.

3.3 CONFORMITY IN INTRA-MODAL RETRIEVAL SIMILARITY

For a query $\mathbf{u}_q$ and gallery sample $\mathbf{u}_j$, both $\in \mathcal{B}$, retrieval ranks the gallery by cosine similarity $s(q,j) = \mathbf{u}_q^\top \mathbf{u}_j$. Each embedding $\mathbf{u}_i$ can be decomposed into its component along $\overline{\boldsymbol{\mu}}$ and an orthogonal component:

$$\mathbf{u}_i = \alpha_i \overline{\boldsymbol{\mu}} + \mathbf{u}_{i,\perp}, \qquad \alpha_i = \mathbf{u}_i^\top \overline{\boldsymbol{\mu}}, \qquad \mathbf{u}_{i,\perp}^\top \overline{\boldsymbol{\mu}} = 0. \tag{3}$$

Since $\overline{\boldsymbol{\mu}}^\top \overline{\boldsymbol{\mu}} = 1$ and the cross terms vanish by orthogonality, $s(q,j) = \mathbf{u}_q^\top \mathbf{u}_j = \alpha_q \alpha_j + \mathbf{u}_{q,\perp}^\top \mathbf{u}_{j,\perp}$. For an arbitrary gallery embedding $\mathbf{u}_j \notin \mathcal{B}$, its conformity with respect to the reference set is $C_{\mathcal{B}}(\mathbf{u}_j) = \frac{1}{M}\sum_k \mathbf{u}_j^\top \mathbf{u}_k = \|\boldsymbol{\mu}\|\,\alpha_j$. Therefore the cosine similarity score can be rewritten as

$$s(q,j) = \underbrace{\frac{\alpha_q}{\|\boldsymbol{\mu}\|}\, C_{\mathcal{B}}(\mathbf{u}_j)}_{\text{conformity component}} + \underbrace{\mathbf{u}_{q,\perp}^\top \mathbf{u}_{j,\perp}}_{\text{orthogonal similarity}}. \tag{4}$$

For a fixed query, $\alpha_q/\|\boldsymbol{\mu}\|$ is constant across gallery samples. Hence, the first term introduces a systematic bias to the retrieval score that is proportional to the gallery sample's conformity with respect to the reference set $C_{\mathcal{B}}(\mathbf{u}_j)$. Depending on the sign of α_q, this contribution can either increase or decrease the similarity score of high-conformity samples.

3.4 RE-CENTERING AND EMERGENCE OF RESIDUAL CONFORMITY

Since the conformity bias term $C_{\mathcal{B}}(\mathbf{u}_j)$ in Eq. 4 is carried entirely by the reference set mean direction (Lemma 1), the most direct embedding-level correction is mean-centering. As shown in Fig. 2-right-(ii), we subtract the mean of the reference set $\boldsymbol{\mu}$ from each embedding $\mathbf{u}_i \in \mathcal{B}$. Since retrieval compares embeddings by cosine similarity, the centered embeddings are then re-normalized to unit norm, $\mathbf{z}_i = \mathrm{Norm}(\mathbf{u}_i - \boldsymbol{\mu})$. Since $\boldsymbol{\mu} = \|\boldsymbol{\mu}\|\,\overline{\boldsymbol{\mu}}$, subtracting the mean only shifts the projection onto $\overline{\boldsymbol{\mu}}$, $\mathbf{u}_i - \boldsymbol{\mu} = (\alpha_i - \|\boldsymbol{\mu}\|)\,\overline{\boldsymbol{\mu}} + \mathbf{u}_{i,\perp}$, which gives the normalization weight and the centered similarity:

$$\mathbf{z}_i = w_i(\mathbf{u}_i - \boldsymbol{\mu}), \qquad w_i = \|\mathbf{u}_i - \boldsymbol{\mu}\|^{-1} = \big(1 + \|\boldsymbol{\mu}\|^2 - 2\|\boldsymbol{\mu}\|\alpha_i\big)^{-1/2}, \tag{5}$$

$$(\mathbf{u}_i - \boldsymbol{\mu})^\top (\mathbf{u}_k - \boldsymbol{\mu}) = \mathbf{u}_{i,\perp}^\top \mathbf{u}_{k,\perp} + \big(\alpha_i - \|\boldsymbol{\mu}\|\big)\big(\alpha_k - \|\boldsymbol{\mu}\|\big). \tag{6}$$

Mean subtraction thus preserves the inner product $\mathbf{u}_{i,\perp}^\top \mathbf{u}_{k,\perp}$ and centers the conformity projections around their reference average, $\|\boldsymbol{\mu}\|$. Fig. 2-right-(ii) shows that it removes the average projection, but not the conformity axis indicated by the dotted line. Samples on the same side of the average retain a positive axial contribution, whereas samples on opposite sides receive a negative one. Before re-normalization, the candidate-dependent part of this contribution for a fixed query q is $(\alpha_q - \|\boldsymbol{\mu}\|)\alpha_j$, up to an additive constant. Although the centered embeddings sum to zero, w_i is strictly increasing in α_i: embeddings with larger conformity projections lie closer to the conformity direction and

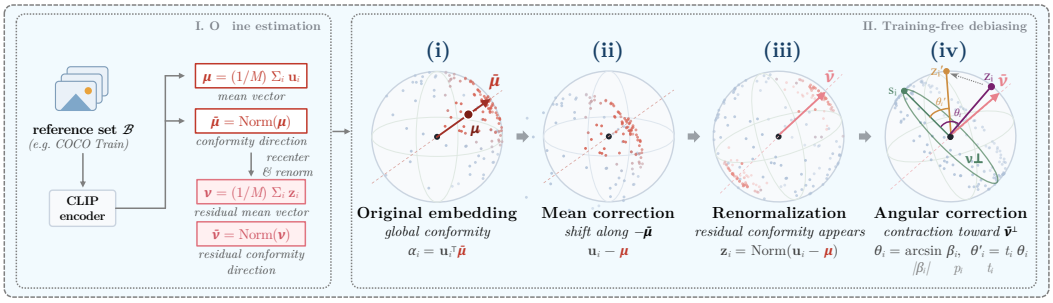


Figure 2: **Conformity-Guided Debiasing (CGD). Left:** CGD estimates its correction statistics once from an external reference set $\mathcal{B}$. Given normalized reference embeddings $\{\mathbf{u}_i\}_{i=1}^M$, it computes the mean $\boldsymbol{\mu}$, the conformity direction $\bar{\boldsymbol{\mu}} = \mathrm{Norm}(\boldsymbol{\mu})$, and, after mean correction and re-normalization $\mathbf{z}_i = \mathrm{Norm}(\mathbf{u}_i - \boldsymbol{\mu})$, the residual mean $\boldsymbol{\nu}$ and the residual conformity direction $\bar{\boldsymbol{\nu}} = \mathrm{Norm}(\boldsymbol{\nu})$. **Right:** At inference, mean correction shifts each embedding by $-\boldsymbol{\mu}$, and re-normalization projects it back onto the unit sphere, where residual conformity remains along $\bar{\boldsymbol{\nu}}$. CGD then contracts the signed angle $\theta_i = \arcsin \beta_i$ toward the subspace orthogonal to $\bar{\boldsymbol{\nu}}$ as $\theta'_i = t_i \theta_i$, where the residual projection $\beta_i = \mathbf{z}_i^\top \bar{\boldsymbol{\nu}}$ sets the correction strength through $t_i = (1 - p_i)/2$, with p_i the reference percentile of $|\beta_i|$. All statistics remain fixed at inference, with no query/gallery fitting.

are amplified more strongly by re-normalization. As shown in Fig. 2-right-(iii), the normalized embeddings do not need to have zero mean, and we define the *residual conformity direction* $\bar{\boldsymbol{\nu}}$ as:

$$\bar{\boldsymbol{\nu}} = \mathrm{Norm}(\boldsymbol{\nu}), \qquad \boldsymbol{\nu} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{z}_i. \tag{7}$$

Additional analysis of the relationship between the global and residual conformity directions is provided in Section A.1. We next introduce a method for attenuating the effect of conformity and residual conformity.

3.5 Conformity-Guided Debiasing (CGD)

We propose *Conformity-Guided Debiasing* (CGD), a training-free post-processing method for intra-modal retrieval that debiases each embedding $\mathbf{u}_i$, with both conformity and residual conformity components attenuated. It is worth noting that $\mathbf{u}_i \notin \mathcal{B}$, and that CGD relies only on the statistics $\boldsymbol{\mu}$ and $\bar{\boldsymbol{\nu}}$ which are estimated once from the reference set $\mathcal{B}$ (Fig. 2, left) and applied to $\mathbf{u}_i$ afterward. For this reason we can state that CGD is *training-free*. It keeps the mean centering and renormalization as its first stage followed by a sample-dependent attenuation along $\bar{\boldsymbol{\nu}}$ that yields the debiased embedding $\mathbf{z}'_i$ used for cosine similarity based retrieval (Fig. 2, right). Given an embedding $\mathbf{u}_i$, CGD defines the progressive mapping

$$\mathbf{u}_i \xmapsto{\text{mean correction}} \mathbf{z}_i \xmapsto{\text{residual attenuation}} \mathbf{z}'_i, \qquad \mathbf{z}_i = \mathrm{Norm}(\mathbf{u}_i - \boldsymbol{\mu}), \qquad \mathbf{u}_i, \mathbf{z}_i, \mathbf{z}'_i \in \mathbb{S}^{d-1}. \tag{8}$$

We now elaborate on the residual attenuation $\mathbf{z}_i \mapsto \mathbf{z}'_i$. To control its strength while keeping the embedding on the unit sphere, we parameterize $\mathbf{z}_i$ by its angle from the subspace orthogonal to $\bar{\boldsymbol{\nu}}$. Let $\beta_i = \mathbf{z}_i^\top \bar{\boldsymbol{\nu}}$ denote the signed residual projection. We decompose $\mathbf{z}_i$ into a unit direction $\mathbf{s}_i$ orthogonal to $\bar{\boldsymbol{\nu}}$ and a signed axial angle θ_i:

$$\mathbf{s}_i = \frac{\mathbf{z}_i - \beta_i \bar{\boldsymbol{\nu}}}{\sqrt{1 - \beta_i^2}}, \qquad \theta_i = \arcsin \beta_i, \qquad \text{so that} \qquad \mathbf{z}_i = \cos\theta_i\, \mathbf{s}_i + \sin\theta_i\, \bar{\boldsymbol{\nu}}, \quad \mathbf{s}_i^\top \bar{\boldsymbol{\nu}} = 0. \tag{9}$$

Here, $\sin\theta_i = \beta_i$ is the signed component of $\mathbf{z}_i$ along the residual axis, while $\cos\theta_i = \sqrt{1 - \beta_i^2}$ is the norm of its orthogonal component, whose direction is $\mathbf{s}_i$. Thus, θ_i is the signed angle of $\mathbf{z}_i$ from the orthogonal subspace, i.e., its displacement along the residual axis. Projecting every embedding onto the subspace orthogonal to $\bar{\boldsymbol{\nu}}$, i.e., replacing $\mathbf{z}_i$ by $\mathbf{s}_i$, which we refer to as *hard projection*, eliminates the residual axial term exactly; however, this complete correction degrades retrieval performance as shown later in Section 4.3, suggesting that the residual direction also contains

Table 1: **Text-to-text retrieval** performance using text encoders of six backbones. We report the task-native primary metric for each benchmark. CGD is applied to Raw and IsoCLIP as a plug-in representation correction. We highlight the best result for each backbone-dataset pair in bold.

	ViT	Method	COCO mAP (%)	Flickr30K mAP (%)	NoCaps mAP (%)	Pascal-50S mAP (%)	CQADupStack nDCG@10 (%)	ESCI nDCG@100 (%)	Quora nDCG@10 (%)	SciFact nDCG@10 (%)	WANDS graded nDCG@10 (%)
CLIP	B/32	Raw	26.2	51.7	35.1	33.6	11.5	6.1	76.2	21.9	54.4
		+ CGD	29.2 (+3.0)	56.1 (+4.4)	39.6 (+4.5)	38.5 (+4.9)	19.5 (+8.0)	13.3 (+7.2)	77.0 (+0.8)	31.5 (+9.6)	62.1 (+7.7)
		IsoCLIP$_{\text{CVPR'26}}$	29.1	56.8	39.6	38.5	23.7	13.5	77.8	35.7	61.7
		+ CGD	**29.8** (+0.7)	**57.8** (+1.0)	**40.3** (+0.7)	**39.3** (+0.8)	**24.7** (+1.0)	**14.5** (+1.0)	**77.9** (+0.1)	**36.4** (+0.7)	**62.7** (+1.0)
	B/16	Raw	26.1	50.7	35.1	33.7	12.1	6.4	76.6	26.3	54.7
		+ CGD	29.2 (+3.1)	55.3 (+4.6)	39.6 (+4.5)	38.4 (+4.7)	20.0 (+7.9)	13.4 (+7.0)	77.3 (+0.7)	36.6 (+10.3)	62.6 (+7.9)
		IsoCLIP$_{\text{CVPR'26}}$	28.9	55.2	39.4	38.2	25.5	14.5	78.1	41.8	63.3
		+ CGD	**29.5** (+0.6)	**56.7** (+1.5)	**40.1** (+0.7)	**38.8** (+0.6)	**26.2** (+0.7)	**15.5** (+1.0)	**78.2** (+0.1)	**42.4** (+0.6)	**64.0** (+0.7)
	L/14	Raw	26.7	52.3	35.7	34.0	14.3	7.5	77.8	23.3	55.8
		+ CGD	30.2 (+3.5)	58.2 (+5.9)	40.5 (+4.8)	39.2 (+5.2)	21.4 (+7.1)	14.4 (+6.9)	78.5 (+0.7)	**37.9** (+14.6)	62.8 (+7.0)
		IsoCLIP$_{\text{CVPR'26}}$	30.1	58.6	40.4	39.3	23.6	14.7	78.7	37.5	63.2
		+ CGD	**30.7** (+0.6)	**60.2** (+1.6)	**41.2** (+0.8)	**40.0** (+0.7)	**23.8** (+0.2)	**15.1** (+0.4)	**78.8** (+0.1)	37.7 (+0.2)	**63.8** (+0.6)
OpenCLIP	B/16	Raw	29.8	57.0	40.0	37.9	24.2	10.2	80.3	34.4	57.8
		+ CGD	31.9 (+2.1)	59.0 (+2.0)	43.1 (+3.1)	41.0 (+3.1)	**26.2** (+2.0)	14.8 (+4.6)	**80.6** (+0.3)	**36.1** (+1.7)	61.1 (+3.3)
		IsoCLIP$_{\text{CVPR'26}}$	31.9	60.8	43.3	41.3	23.9	17.1	79.5	34.7	63.3
		+ CGD	**32.2** (+0.3)	**61.4** (+0.6)	**43.6** (+0.3)	**41.5** (+0.2)	24.4 (+0.5)	**17.7** (+0.6)	79.8 (+0.3)	34.8 (+0.1)	**63.8** (+0.5)
SigLIP	B/16	Raw	27.0	56.4	38.6	35.8	19.4	9.6	80.5	28.5	54.5
		+ CGD	**31.5** (+4.5)	**61.2** (+4.8)	**44.5** (+5.9)	**41.6** (+5.8)	**25.2** (+5.8)	16.9 (+7.3)	**81.2** (+0.7)	**41.4** (+12.9)	61.3 (+6.8)
		IsoCLIP$_{\text{CVPR'26}}$	29.1	57.3	42.0	39.3	23.4	16.0	80.8	37.0	60.6
		+ CGD	30.9 (+1.8)	58.2 (+0.9)	43.3 (+1.3)	40.7 (+1.4)	23.7 (+0.3)	**17.4** (+1.4)	80.9 (+0.1)	37.3 (+0.3)	**61.9** (+1.3)
SigLIP2	B/16	Raw	29.2	56.4	38.8	36.1	12.4	12.4	79.5	24.7	56.8
		+ CGD	33.4 (+4.2)	**62.1** (+5.7)	44.9 (+6.1)	42.8 (+6.7)	**16.4** (+4.0)	17.7 (+5.3)	**79.9** (+0.4)	**30.5** (+5.8)	62.0 (+5.2)
		IsoCLIP$_{\text{CVPR'26}}$	32.4	58.3	44.0	42.0	15.6	18.3	79.4	26.9	62.8
		+ CGD	**33.6** (+1.2)	61.4 (+3.1)	**45.0** (+1.0)	**43.1** (+1.1)	15.9 (+0.3)	**19.0** (+0.7)	79.4 (+0.0)	27.2 (+0.3)	**63.8** (+1.0)

retrieval-relevant information. CGD therefore applies a partial, sample-dependent correction whose strength is determined by $|\beta_i|$. For each embedding $\mathbf{z}_i$, let

$$p_i = F_{\text{ref}}(|\beta_i|) \in [0, 1], \qquad t_i = \frac{1 - p_i}{2}, \tag{10}$$

where p_i is the empirical percentile of $|\beta_i|$ in the reference set. The same reference distribution is used for queries and gallery items.

CGD uses spherical linear interpolation (SLERP) Shoemake (1985) to move $\mathbf{z}_i$ toward its orthogonal projection $\mathbf{s}_i$. Letting θ_i' denote the debiased residual angle, the debiased embedding is

$$bz_i' = \text{SLERP}(\mathbf{s}_i, \theta_i; t_i) = \cos\theta_i' \, \mathbf{s}_i + \sin\theta_i' \, \overline{\boldsymbol{\nu}}, \qquad \theta_i' = t_i \theta_i. \tag{11}$$

The interpolation parameter $t_i \in [0, \frac{1}{2}]$ controls the degree of residual attenuation: $t_i = 0$ yields the hard projection $\mathbf{s}_i$, while larger values retain a greater fraction of the original residual angle. Intermediate values partially attenuate the residual component while preserving the orthogonal direction $\mathbf{s}_i$ and maintaining unit norm. Since t_i decreases with $|\beta_i|$, embeddings with stronger residual projections undergo greater angular contraction, whereas those with weaker projections undergo a milder correction. Moreover, $t_i \leq \frac{1}{2}$ ensures that at least half of the residual angle is removed from every embedding. The complete CGD algorithm and implementation details are provided in Section A.2. Additional analysis of pairwise similarities is provided in Section A.3.

4 Experiments

4.1 Experimental Settings

We evaluate CGD on text-to-text (T2T) retrieval, image-to-image (I2I) retrieval, and image-only few-shot classification. CGD is applied both to Raw representations and on top of existing intra-modal corrections, including IsoCLIP (Magistri et al., 2026) and PCA← (Herzog & Wang, 2026). We provide real-space visualizations of CGD across these representations in Section A.4.

Datasets and metrics. For T2T retrieval, we evaluate on nine benchmarks covering caption retrieval, product search, duplicate-question retrieval, and scientific evidence retrieval: COCO Text Lin et al. (2014), Flickr30K Text Young et al. (2014), NoCaps Text Agrawal et al. (2019), PASCAL-50S Vedantam et al. (2015), ESCI Reddy et al. (2022); Hou et al. (2026), WANDS Chen et al. (2022), CQADupStack-English Hoogeveen et al. (2015), SciFact Wadden et al. (2020), and QuoraRetrieval Thakur et al. (2021). We follow the benchmark-specific evaluation metrics. For I2I

Table 2: **Image-to-image retrieval** performance using vision encoders of six backbones. CGD is applied to Raw, IsoCLIP and PCA← as a plug-in representation correction.

	ViT	Method	Cal.	CUB	ROxf.	RPar.	Cars	Pets	Flw.	Air.	DTD	Euro.	Food	SUN	UCF	Avg.
CLIP	B/32	Raw	77.1	22.9	42.6	67.9	24.6	30.5	62.0	14.5	28.1	47.9	32.3	34.3	47.1	40.9
		+ CGD	78.8 (+1.7)	23.5 (+0.6)	44.7 (+2.1)	69.7 (+1.8)	25.6 (+1.0)	33.2 (+2.7)	63.6 (+1.6)	14.6 (+0.1)	29.5 (+1.4)	50.4 (+2.5)	34.4 (+2.1)	34.9 (+0.6)	47.3 (+0.2)	42.3 (+1.4)
		IsoCLIP$_{\text{CVPR'26}}$	80.8	27.0	47.2	73.8	30.0	40.8	66.5	14.9	30.9	51.5	38.0	36.4	48.4	45.1
		+ CGD	81.7 (+0.9)	27.4 (+0.4)	47.5 (+0.3)	74.0 (+0.2)	30.8 (+0.8)	42.2 (+1.4)	67.2 (+0.7)	15.0 (+0.1)	31.2 (+0.3)	52.2 (+0.7)	39.3 (+1.3)	37.1 (+0.7)	47.9 (−0.5)	45.7 (+0.6)
		PCA←$_{\text{CVPR'26}}$	83.1	30.3	47.8	74.5	34.6	47.5	70.3	16.0	**33.9**	53.8	43.0	38.2	53.1	48.2
		+ CGD	**83.2** (+0.1)	**31.4** (+1.1)	**49.1** (+1.3)	**75.2** (+0.7)	**34.7** (+0.1)	**48.1** (+0.6)	**70.4** (+0.1)	**16.2** (+0.2)	**33.9**	**54.0** (+0.2)	**43.3** (+0.3)	**40.0** (+1.8)	**53.3** (+0.2)	**48.7** (+0.5)
	B/16	Raw	80.6	31.6	46.6	75.3	31.0	36.3	70.8	19.0	30.7	51.2	42.8	35.9	49.8	46.3
		+ CGD	81.9 (+1.3)	34.6 (+3.0)	47.7 (+1.1)	77.0 (+1.7)	32.3 (+1.3)	39.0 (+2.7)	72.7 (+1.9)	19.3 (+0.3)	32.3 (+1.6)	52.1 (+0.9)	44.2 (+1.4)	37.9 (+2.0)	51.2 (+1.4)	47.9 (+1.6)
		IsoCLIP$_{\text{CVPR'26}}$	85.0	38.6	51.8	82.0	41.2	50.7	77.4	20.5	36.0	55.6	**53.5**	38.5	55.4	52.8
		+ CGD	**85.5** (+0.5)	**40.8** (+2.2)	**52.5** (+0.7)	**82.3** (+0.3)	**41.4** (+0.2)	**51.6** (+0.9)	**77.8** (+0.4)	**20.7** (+0.2)	**36.3** (+0.3)	**55.7** (+0.1)	**53.5**	40.1 (+1.6)	**56.2** (+0.8)	**53.4** (+0.6)
		PCA←$_{\text{CVPR'26}}$	84.4	40.2	50.5	79.9	40.1	47.1	77.0	20.3	35.6	53.4	51.0	40.3	54.8	51.9
		+ CGD	85.2 (+0.8)	40.4 (+0.2)	51.1 (+0.6)	80.5 (+0.6)	40.5 (+0.4)	48.0 (+0.9)	77.4 (+0.4)	20.4 (+0.1)	**36.3** (+0.7)	54.0 (+0.6)	51.6 (+0.6)	**41.3** (+1.0)	55.8 (+1.0)	52.5 (+0.6)
	L/14	Raw	83.2	43.0	57.5	76.9	43.3	47.3	84.0	25.8	34.1	59.0	53.0	39.1	60.0	54.3
		+ CGD	86.6 (+3.4)	45.1 (+2.1)	60.2 (+2.7)	78.7 (+1.8)	46.6 (+3.3)	51.5 (+4.2)	86.4 (+2.4)	26.6 (+0.8)	36.8 (+2.7)	60.3 (+1.3)	57.5 (+4.5)	41.6 (+2.5)	61.9 (+1.9)	56.9 (+2.6)
		IsoCLIP$_{\text{CVPR'26}}$	87.0	52.2	66.4	81.4	56.4	63.5	88.2	28.2	39.0	61.6	62.9	41.0	61.9	60.7
		+ CGD	87.9 (+0.9)	53.1 (+0.9)	**67.0** (+0.6)	81.9 (+0.5)	56.9 (+0.5)	**64.0** (+0.5)	88.5 (+0.3)	28.4 (+0.2)	39.3 (+0.3)	61.0 (−0.6)	63.7 (+0.8)	41.5 (+0.5)	62.5 (+0.6)	61.2 (+0.5)
		PCA←$_{\text{CVPR'26}}$	89.4	54.3	66.5	83.2	56.7	62.7	88.7	28.6	39.9	62.7	64.3	44.7	66.7	62.2
		+ CGD	**89.9** (+0.5)	**54.7** (+0.4)	**67.0** (+0.5)	**83.9** (+0.7)	**57.1** (+0.4)	63.4 (+0.7)	**88.9** (+0.2)	**28.8** (+0.2)	**40.0** (+0.1)	**62.9** (+0.2)	**64.8** (+0.5)	**46.3** (+1.6)	**67.1** (+0.4)	**62.7** (+0.5)
OpenCLIP	B/16	Raw	85.7	42.8	65.3	83.2	55.8	50.4	84.6	23.1	39.9	57.8	51.1	39.5	52.9	56.3
		+ CGD	86.5 (+0.8)	43.7 (+0.9)	65.6 (+0.3)	84.1 (+0.9)	56.9 (+1.1)	52.6 (+2.2)	84.6	23.3 (+0.2)	40.4 (+0.5)	58.3 (+0.5)	51.6 (+0.5)	40.2 (+0.7)	53.9 (+1.0)	57.1 (+0.8)
		IsoCLIP$_{\text{CVPR'26}}$	87.6	45.9	67.3	85.0	60.7	57.8	85.8	23.5	42.5	58.6	54.7	39.3	53.4	58.6
		+ CGD	87.7 (+0.1)	46.3 (+0.4)	67.9 (+0.6)	**86.4** (+1.4)	60.9 (+0.2)	58.3 (+0.5)	85.8	23.6 (+0.1)	42.6 (+0.1)	59.2 (+0.6)	54.3 (−0.4)	40.2 (+0.9)	54.2 (+0.8)	59.0 (+0.4)
		PCA←$_{\text{CVPR'26}}$	89.5	50.1	69.2	85.5	65.3	62.9	**86.8**	25.1	**44.8**	59.4	57.1	42.7	56.5	61.1
		+ CGD	**89.7** (+0.2)	**50.2** (+0.1)	**69.7** (+0.5)	85.8 (+0.3)	**65.5** (+0.2)	**63.1** (+0.2)	86.7 (−0.1)	**25.3** (+0.2)	44.7 (−0.1)	**59.8** (+0.4)	**57.4** (+0.3)	**43.1** (+0.4)	**57.0** (+0.5)	**61.4** (+0.3)
SigLIP	B/16	Raw	87.2	39.2	50.6	73.1	65.7	56.4	87.5	37.9	39.8	53.3	56.3	42.8	56.9	57.4
		+ CGD	89.5 (+2.3)	40.7 (+1.5)	51.7 (+1.1)	76.5 (+3.4)	68.7 (+3.0)	59.7 (+3.3)	88.7 (+1.2)	38.3 (+0.4)	41.0 (+1.2)	53.4 (+0.1)	58.0 (+1.7)	43.8 (+1.0)	57.5 (+0.6)	59.0 (+1.6)
		IsoCLIP$_{\text{CVPR'26}}$	89.8	41.0	50.0	75.0	67.2	60.7	85.0	35.2	42.2	54.4	59.4	42.7	58.1	58.5
		+ CGD	91.1 (+1.3)	41.7 (+0.7)	50.5 (+0.5)	76.7 (+1.7)	68.6 (+1.4)	62.8 (+2.1)	85.8 (+0.8)	35.8 (+0.6)	42.7 (+0.5)	**54.5** (+0.1)	60.1 (+0.7)	44.1 (+1.4)	58.8 (+0.7)	59.5 (+1.0)
		PCA←$_{\text{CVPR'26}}$	91.2	47.0	57.9	78.4	77.2	68.5	92.0	40.9	44.2	54.2	61.8	46.9	60.3	63.1
		+ CGD	**91.9** (+0.7)	**47.1** (+0.1)	**58.1** (+0.2)	**78.7** (+0.3)	**78.9** (+1.7)	**71.9** (+3.4)	**92.5** (+0.5)	**41.7** (+0.8)	**44.4** (+0.2)	53.9 (−0.3)	**62.8** (+1.0)	**47.2** (+0.3)	**61.2** (+0.9)	**63.9** (+0.8)
SigLIP2	B/16	Raw	89.2	38.1	53.2	76.5	70.8	56.6	89.3	41.8	39.0	50.8	59.2	43.6	59.2	59.0
		+ CGD	90.9 (+1.7)	39.4 (+1.3)	54.3 (+1.1)	77.7 (+1.2)	73.1 (+2.3)	59.7 (+3.1)	90.4 (+1.1)	42.4 (+0.6)	39.9 (+0.9)	50.9 (+0.1)	60.1 (+0.9)	44.1 (+0.5)	60.4 (+1.2)	60.3 (+1.3)
		IsoCLIP$_{\text{CVPR'26}}$	93.1	41.4	54.2	77.9	74.5	64.0	87.2	40.9	43.8	53.2	63.6	46.9	61.8	61.7
		+ CGD	**93.8** (+0.7)	41.9 (+0.5)	54.6 (+0.4)	78.7 (+0.8)	74.5	65.4 (+1.4)	87.8 (+0.6)	41.2 (+0.3)	43.9 (+0.1)	**53.3** (+0.1)	64.0 (+0.4)	47.5 (+0.6)	62.0 (+0.2)	62.2 (+0.5)
		PCA←$_{\text{CVPR'26}}$	93.0	47.3	59.4	78.6	80.2	67.8	93.2	46.3	**44.1**	51.1	65.3	**48.9**	63.0	64.5
		+ CGD	**93.8** (+0.8)	**47.4** (+0.1)	**60.4** (+1.0)	**79.5** (+0.9)	**82.1** (+1.9)	**72.4** (+4.6)	**93.7** (+0.5)	**48.9** (+2.6)	**44.1**	51.4 (+0.3)	**66.5** (+1.2)	**48.9**	**64.2** (+1.2)	**65.6** (+1.1)

retrieval, we follow the 13-dataset protocol of IsoCLIP and report mAP; for image-only few-shot classification, we follow the 11-dataset setting of PCA← and report accuracy. Full dataset and evaluation details are provided in the Appendix Section B.1.

Backbones and baselines. We evaluate six frozen VLM backbones: CLIP ViT-B/32, ViT-B/16, ViT-L/14, OpenCLIP ViT-B/16, SigLIP ViT-B/16, and SigLIP2 ViT-B/16. For T2T, we compare Raw and IsoCLIP with and without CGD; for I2I, we additionally include PCA←, which does not define a T2T formulation.

CGD implementation details. We use COCO training images and captions as the default reference sets for I2I and T2T, respectively. CGD is fitted once and applied unchanged to downstream data. Using ImageNet images or Flickr30K captions as alternative reference sets gives similar improvements (Section B.2). CGD adds negligible overhead: fitting 82,783 embeddings takes 1.9 ms, and processing one million embeddings takes 0.033 s.

4.2 Main comparison

Text-to-Text (T2T) Retrieval. Table 1 reports T2T retrieval results across nine benchmarks. CGD improves Raw representations on all 54 backbone-dataset pairs, with gains up to +14.6 on SciFact for CLIP ViT-L/14. Applied on top of IsoCLIP, CGD also yields uniformly positive gains, including +3.1 on Flickr30K with SigLIP2 B/16, demonstrating its complementarity to existing spectral correction.

Image-to-Image (I2I) Retrieval. Table 2 reports I2I retrieval results across 13 benchmarks. CGD consistently improves the Raw representations. On CLIP ViT-L/14, CGD improves the average mAP from 54.3 to 56.9 (+2.6), with gains of +4.5 on Food101 and +4.2 on Oxford Pets. Notably, CGD provides further improvements on top of both IsoCLIP and PCA←. Results show that conformity correction is complementary to both spectral and subspace-based corrections for I2I retrieval.

Image-Only Few-Shot Classification. Following PCA←, we further evaluate CGD on image-only few-shot classification using two backbones and no text prompts. We consider Prototype classification based on class-mean embeddings and LDA with a shared within-class covariance, and report average accuracy over 11 datasets for 1, 2, 4, 8, and 16 shots. Table 3 shows that CGD consistently improves

Table 3: **Image-only few-shot classification** with the evaluation protocol of PCA←. We report average classification accuracy (%) over 11 datasets using only image-to-image similarities without text prompts. Results of +CGD variants are averaged over three random seeds. LDA cannot estimate covariance for the 1-shot setting; the reported 2.3% is the average prevalence of the default class.

	Classifier	Feature	1	2	4	8	16
CLIP B/16	Prototype	Raw	43.4	55.3	63.8	69.6	73.5
		Raw + **CGD**	45.7 (+2.3)	56.3 (+1.0)	65.4 (+1.6)	70.6 (+1.0)	74.2 (+0.7)
		IsoCLIP	50.0	60.7	69.2	74.0	76.9
		IsoCLIP + **CGD**	50.6 (+0.6)	61.0 (+0.3)	69.3 (+0.1)	74.1 (+0.1)	77.0 (+0.1)
		PCA←	50.7	62.5	69.7	74.7	77.5
		PCA← + **CGD**	**52.5** (+1.8)	**62.6** (+0.1)	**70.9** (+1.2)	**75.4** (+0.7)	**78.1** (+0.6)
	LDA	Raw	2.3	60.0	69.8	76.1	79.5
		Raw + **CGD**	2.3	61.0 (+1.0)	71.2 (+1.4)	76.7 (+0.6)	80.3 (+0.8)
		IsoCLIP	2.3	61.9	71.2	76.4	79.8
		IsoCLIP + **CGD**	2.3	62.2 (+0.3)	71.3 (+0.1)	76.4	79.8
		PCA←	2.3	62.5	70.8	76.5	79.6
		PCA← + **CGD**	**2.3**	**63.6** (+1.1)	**72.3** (+1.5)	**77.2** (+0.7)	**80.4** (+0.8)

	Classifier	Feature	1	2	4	8	16
SigLIP2 B/16	Prototype	Raw	58.0	69.7	77.0	80.8	83.0
		Raw + **CGD**	59.4 (+1.4)	70.2 (+0.5)	77.3 (+0.3)	81.3 (+0.5)	83.4 (+0.4)
		IsoCLIP	59.8	70.2	77.6	81.6	83.6
		IsoCLIP + **CGD**	60.5 (+0.7)	70.4 (+0.2)	77.8 (+0.2)	81.7 (+0.1)	83.7 (+0.1)
		PCA←	63.4	73.2	**79.4**	82.6	84.5
		PCA← + **CGD**	**63.8** (+0.4)	**73.3** (+0.1)	79.2 (−0.2)	**82.8** (+0.2)	**84.6** (+0.1)
	LDA	Raw	2.3	73.2	80.5	84.5	86.5
		Raw + **CGD**	2.3	73.4 (+0.2)	80.8 (+0.3)	84.9 (+0.4)	86.9 (+0.4)
		IsoCLIP	2.3	73.3	80.8	84.6	86.7
		IsoCLIP + **CGD**	2.3	73.5 (+0.2)	80.8	84.6	86.7
		PCA←	2.3	74.5	81.0	84.7	86.5
		PCA← + **CGD**	**2.3**	**74.6** (+0.1)	**81.3** (+0.3)	**84.9** (+0.2)	**86.9** (+0.4)

classification, with larger gains in low-shot regimes where shared non-discriminative components have a stronger impact. As the number of shots increases, this benefit becomes smaller.

4.3 Ablations

We study three aspects of CGD: the design choices behind the correction, the effect of the reference-set source, and the number of reference samples required. Additional experiments on bidirectional residual correction, sensitivity to angular retention, and iterative CGD are provided in Section C.

Design Choices in CGD. In Table 4, we compare several correction strategies motivated by the analysis to clarify the design choices leading to CGD. We evaluate CLIP ViT-B/32 on three I2I benchmarks and two T2T benchmarks. We compare several alternatives motivated by our analysis: subtraction of the conformity component term from the similarity score in Eq. (4), embedding-level hard projection and mean centering ($1\times$) discussed in Section 3.5 and Section 3.4 respectively. We additionally evaluate repeated centering (indicated as $2\times$ in the table), in which we recenter using the residual mean $\boldsymbol{\nu}$ from Eq. (7) after the initial centering followed by L_2 normalization. Full CGD instead retains the initial mean-centering and L_2 normalization step, followed by adaptive partial angular attenuation along the residual conformity direction $\overline{\boldsymbol{\nu}}$ from Eq. (7).

Table 4: **Design choices in CGD.** Comparison of alternative conformity correction strategies on selected I2I and T2T retrieval benchmarks. Results are reported in mAP; green values indicate the improvement of CGD over Raw.

	Dataset	Raw	Direct removal		Centering		CGD (imp. over Raw)
			score-level Direct subtraction	*embedding-level* Hard projection	$1\times$	$2\times$	
I2I	EuroSAT	47.93	45.02	45.02	48.89	48.87	**50.42** (+2.49)
	Oxford Pets	30.53	30.30	30.30	31.84	31.77	**33.19** (+2.66)
	Food101	32.32	30.81	30.80	33.21	33.08	**34.41** (+2.09)
	Avg.	36.93	35.38	35.37	37.98	37.91	**39.34** (+2.41)
T2T	Flickr30K	51.68	46.25	46.25	50.82	50.59	**56.08** (+4.40)
	NoCaps	35.07	36.29	36.29	37.66	37.60	**39.63** (+4.56)
	Avg.	43.38	41.27	41.27	44.24	44.10	**47.86** (+4.48)

As shown in Table 4, our CGD achieves the strongest performance on every evaluated dataset. These results confirm that global and residual conformity correction are two complementary steps toward a less biased and more discriminative representation.

Impact of reference set source. Using CLIP ViT-B/32, we compare four reference-set choices with 1,000 samples: the default COCO training set, an *in-distribution* held-out split, a *mixed-distribution* set, and an *out-of-distribution* set that excludes the downstream benchmark. As shown in Table 5, in-distribution references generally give the largest gains, while out-of-distribution references still improve performance on all six benchmarks. This indicates that CGD benefits from distribution-matched references when available, but does not require access to the target distribution. We therefore

use COCO as the fixed default reference set throughout the main experiments. Full construction details are given in Appendix B.3.

Impact of reference set Size. We study the effect of the COCO reference-bank size in Fig. 3 on two I2I retrieval benchmarks (Oxford-IIIT Pets and Stanford Cars) and two T2T retrieval benchmarks (COCO and NoCaps). CGD already provides positive gains with as few as 5-50 reference samples, and performance improves rapidly as the bank size increases. With around 2K reference samples, the gains are already close to those obtained with the full COCO bank. This trend is consistent across different backbones and both I2I and T2T retrieval, showing that CGD does not require a large reference set to estimate reliable conformity statistics.

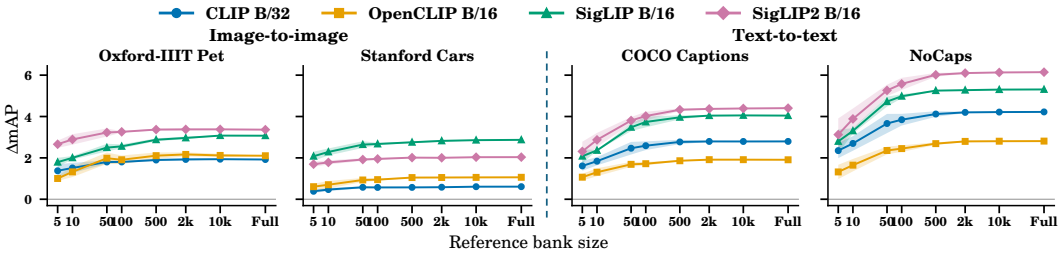


Figure 3: **Robustness to reference bank size.** mAP gain of CGD over Raw when varying the size of the COCO reference bank on two I2I and two T2T retrieval benchmarks. CGD provides positive gains with only a few reference samples and largely stabilizes at around 2K samples across backbones. Results are averaged over three random subsampling seeds.

Table 5: **Effect of reference-bank source.** mAP gain of CGD over Raw using the default COCO, in-distribution, mixed-distribution, and out-of-distribution reference data.

I2I				
Dataset	Default	In-dist.	Mixed-dist.	Out-of-dist.
DTD	+1.17	**+3.45**	+2.08	+1.54
Food101	+0.77	**+2.67**	+1.10	+0.70
Caltech	+0.96	**+1.85**	+1.70	+1.61

T2T				
Dataset	Default	In-dist.	Mixed-dist.	Out-of-dist.
COCO	+3.03	**+3.04**	+3.00	+2.00
Flickr	+5.11	+6.79	**+7.07**	+0.83
NoCaps	+4.88	**+5.25**	+5.15	+3.39

5 Limitations

Although CGD consistently improves intra-modal retrieval and few-shot classification, several limitations remain. First, the method depends on a reference set from which the conformity statistics are estimated. While our experiments show that CGD remains effective with out-of-distribution reference data, distribution-matched references generally provide larger gains. The sensitivity of the method to the size, composition, and domain of the reference set therefore deserves further investigation. Second, CGD models conformity through a single dominant direction. This assumption may be insufficient for representations characterized by multiple conformity factors or multimodal geometric structures. Extending the analysis to several directions, for example through a structured spectral or subspace-based formulation, could provide a more expressive correction. At the same time, a single dominant direction ensures a lightweight computation, which could become heavier in the case of spectral analysis.

6 Conclusion

We introduced Conformity-Guided Debiasing, a training-free and computationally lightweight method for correcting conformity-related biases in normalized representations. Our analysis shows that conformity can be characterized geometrically through a preferred direction in the unit hypersphere. Mean-centering removes the global conformity component, but renormalization gives rise to a residual conformity effect that cannot be eliminated by centering alone. CGD addresses this residual effect through sample-dependent angular attenuation. The method requires no retraining of the encoder, and can be fitted once on a reference set and applied unchanged to downstream data.

AI USE STATEMENT

Generative AI tools were used to assist with polishing the manuscript. All AI-assisted content was carefully reviewed, verified, and revised by the authors as appropriate. The authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

All datasets used in this work are publicly available. We provide the complete formulation and implementation procedure of CGD in the main paper and Appendix Section A, and document the datasets, evaluation protocols, reference-set construction, and additional experimental settings in Appendix Section B. Additional ablations and sensitivity analyses are provided in Appendix Section C. Upon publication, we will release the complete codebase together with all evaluation scripts, configurations, and cached intermediate files used to produce the reported results, including cached model representations, to facilitate direct reproduction of our experiments.

REFERENCES

Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 8947–8956. IEEE, 2019. doi: 10.1109/ICCV.2019.00904. URL https://doi.org/10.1109/ICCV.2019.00904.

Josh Beal, Eric Kim, Jinfeng Rao, Rex Wu, Dmitry Kislyuk, and Charles Rosenberg. Pinclip: Large-scale foundational multimodal representation at pinterest. *CoRR*, abs/2603.03544, 2026. doi: 10.48550/ARXIV.2603.03544. URL https://doi.org/10.48550/arXiv.2603.03544.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, volume 8694 of *Lecture Notes in Computer Science*, pp. 446–461. Springer, 2014. doi: 10.1007/978-3-319-10599-4_29. URL https://doi.org/10.1007/978-3-319-10599-4_29.

Yan Chen, Shujian Liu, Zheng Liu, Weiyi Sun, Linas Baltrunas, and Benjamin Schroeder. WANDS: dataset for product search relevance assessment. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (eds.), *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I*, volume 13185 of *Lecture Notes in Computer Science*, pp. 128–141. Springer, 2022. doi: 10.1007/978-3-030-99736-6_9. URL https://doi.org/10.1007/978-3-030-99736-6_9.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 3606–3613. IEEE Computer Society, 2014. doi: 10.1109/CVPR.2014.461. URL https://doi.org/10.1109/CVPR.2014.461.

Li Fei-Fei, Robert Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, 2007. doi: 10.1016/J.CVIU.2005.09.012. URL https://doi.org/10.1016/j.cviu.2005.09.012.

Yossi Gandelsman, Alexei Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. In *International conference on learning representations*, volume 2024, pp. 18395–18416, 2024.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242. URL https://doi.org/10.1109/JSTARS.2019.2918242.

Jonas Herzog and Yue Wang. Reevaluating the intra-modal misalignment hypothesis in clip. In *Computer Vision and Pattern Recognition Conference*, 2026.

Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. Cqadupstack: A benchmark data set for community question-answering research. In Laurence Anthony F. Park and Sarvnaz Karimi (eds.), *Proceedings of the 20th Australasian Document Computing Symposium, ADCS 2015, Parramatta, NSW, Australia, December 8-9, 2015*, pp. 3:1–3:8. ACM, 2015. doi: 10.1145/2838931.2838934. URL https://doi.org/10.1145/2838931.2838934.

Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian J. McAuley. Bridging language and items for retrieval and recommendation: Benchmarking llms as semantic encoders. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 3251–3265. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.147. URL https://doi.org/10.18653/v1/2026.acl-long.147.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pp. 554–561. IEEE Computer Society, 2013. doi: 10.1109/ICCVW.2013.77. URL https://doi.org/10.1109/ICCVW.2013.77.

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of CLIP. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33999–34019. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/levi25b.html.

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. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/702f4db7543a7432431df588d57bc7c9-Abstract-Conference.html.

Zihan Liang, Yufei Ma, Zhipeng Qian, Huangyu Dai, Zihan Wang, Ben Chen, Chenyi Lei, Yuqing Ding, and Han Li. Uniecs: Unified multimodal e-commerce search framework with gated cross-modal fusion. In Meeyoung Cha, Chanyoung Park, Noseong Park, Carl Yang, Senjuti Basu Roy, Jessie Li, Jaap Kamps, Kijung Shin, Bryan Hooi, and Lifang He (eds.), *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM 2025, Seoul, Republic of Korea, November 10-14, 2025*, pp. 1788–1797. ACM, 2025. doi: 10.1145/3746252.3761170. URL https://doi.org/10.1145/3746252.3761170.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Zengrong Lin, Zheng Wang, Tianwen Qian, Pan Mu, Sixian Chan, and Cong Bai. Neighborretr: Balancing hub centrality in cross-modal retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 9263–9273. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00865. URL https://openaccess.thecvf.com/content/CVPR2025/html/Lin_NeighborRetr_Balancing_Hub_Centrality_in_Cross-Modal_Retrieval_CVPR_2025_paper.html.

Simone Magistri, Dipam Goswami, Marco Mistretta, Bartłomiej Twardowski, Joost van de Weijer, and Andrew D Bagdanov. Isoclip: Decomposing clip projectors for efficient intra-modal alignment. In *Computer Vision and Pattern Recognition Conference*, 2026.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. URL http://arxiv.org/abs/1306.5151.

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.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, pp. 722–729. IEEE Computer Society, 2008. doi: 10.1109/ICVGIP.2008.47. URL https://doi.org/10.1109/ICVGIP.2008.47.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pp. 3498–3505. IEEE Computer Society, 2012. doi: 10.1109/CVPR.2012.6248092. URL https://doi.org/10.1109/CVPR.2012.6248092.

Filip Radenovic, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 5706–5715. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00598. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Radenovic_Revisiting_Oxford_and_CVPR_2018_paper.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. Shopping queries dataset: A large-scale ESCI benchmark for improving product search. *CoRR*, abs/2206.06588, 2022. doi: 10.48550/ARXIV.2206.06588. URL https://doi.org/10.48550/arXiv.2206.06588.

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.

Ken Shoemake. Animating rotation with quaternion curves. In Pat Cole, Robert Heilman, and Brian A. Barsky (eds.), *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1985, San Francisco, California, USA, July 22-26, 1985*, pp. 245–254. ACM, 1985. doi: 10.1145/325334.325242. URL https://doi.org/10.1145/325334.325242.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. URL http://arxiv.org/abs/1212.0402.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 4566–4575. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://doi.org/10.1109/CVPR.2015.7299087.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.609. URL https://doi.org/10.18653/v1/2020.emnlp-main.609.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. Caltech-ucsd birds-200-2011. https://doi.org/10.21227/fg6w-vh29, June 2025. URL https://doi.org/10.21227/fg6w-vh29. Accessed on YYYY-MM-DD.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pp. 3485–3492. IEEE Computer Society, 2010. doi: 10.1109/CVPR.2010.5539970. URL https://doi.org/10.1109/CVPR.2010.5539970.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014. doi: 10.1162/TACL_A_00166. URL https://doi.org/10.1162/tacl_a_00166.

Xinliang Zhu, Sheng-Wei Huang, Han Ding, Jinyu Yang, Kelvin Chen, Tao Zhou, Tal Neiman, Ouye Xie, Son Tran, Benjamin Z. Yao, Douglas Gray, Anuj Bindal, and Arnab Dhua. Bringing multimodality to amazon visual search system. In Ricardo Baeza-Yates and Francesco Bonchi (eds.), *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pp. 6390–6399. ACM, 2024. doi: 10.1145/3637528.3671640. URL https://doi.org/10.1145/3637528.3671640.

Appendix organization. The appendix is organized as follows:

- **Section A: Additional Analysis of CGD.** We analyze the relationship between global and residual conformity (Section A.1), provide the detailed CGD implementation (Section A.2), study its effect on pairwise similarity (Section A.3), and visualize the correction in real embedding spaces (Section A.4).
- **Section B: Experimental Details.** We provide the dataset and evaluation protocols (Section B.1), evaluate alternative reference sets (Section B.2), and describe the construction of the reference-bank source ablations (Section B.3).
- **Section C: Additional Ablations.** We further study the residual correction design, sensitivity to angular retention, and iterative CGD.

A Additional Analysis of CGD

A.1 Global and Residual Conformity Directions

We further analyze the relationship between the global conformity direction and the residual conformity direction that emerges after mean correction and re-normalization. All experiments use CLIP ViT-B/32 and five random seeds, with $M = 1{,}000$ reference samples for each direction estimate.

Recall from Sec. 3.2 that, for a normalized reference set $\mathcal{B} = \{\mathbf{u}_1, \ldots, \mathbf{u}_M\}$, its mean and corresponding conformity direction are

$$\boldsymbol{\mu} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{u}_i, \qquad \overline{\boldsymbol{\mu}} = \mathrm{Norm}(\boldsymbol{\mu}). \tag{12}$$

After mean correction and re-normalization, we obtain

$$\mathbf{z}_i = \mathrm{Norm}(\mathbf{u}_i - \boldsymbol{\mu})\,. \tag{13}$$

The mean of these corrected representations and its corresponding direction are

$$\boldsymbol{\nu} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{z}_i, \qquad \overline{\boldsymbol{\nu}} = \mathrm{Norm}(\boldsymbol{\nu}), \tag{14}$$

where $\overline{\boldsymbol{\nu}}$ is the *residual conformity direction*.

For the analyses below, we consider both the default COCO reference sets and alternative reference sets with different data distributions. For I2I, *Self-val mix-1K* denotes a 1,000-image reference set constructed by approximately balancing held-out validation images from the I2I retrieval datasets, with the reference images disjoint from the corresponding query/gallery splits. For T2T, *Medical+Finance-1K* contains 500 medical queries from NFCorpus and 500 financial queries from FiQA. The detailed construction of these reference sets is provided in Section B.

Analysis 1: Is residual conformity simply a re-emergence of global conformity?

To answer this question, we measure the alignment between the global conformity direction $\overline{\boldsymbol{\mu}}$ and the residual conformity direction $\overline{\boldsymbol{\nu}}$ estimated from the same reference set. As shown in Table 6, across both image and text representations, their cosine similarity ranges from 0.475 to 0.600, corresponding to directional angles of approximately $53°$–$62°$. Thus, the residual conformity direction remains positively aligned with the global conformity direction, but is far from identical to it.

Analysis 2: How reference-dependent are the global and residual conformity directions?

We next examine how the two directions change when the external reference set is replaced. For I2I, we compare directions estimated from COCO images with those estimated from a balanced mixture of held-out validation images. For T2T, we compare COCO captions with a cross-domain mixture of medical and financial text. As shown in Table 7, the global conformity direction is substantially more stable across reference sets than the residual conformity direction. For I2I, the cross-set cosine

Table 6: **Relationship between the global and residual conformity directions.** We report the cosine similarity between $\overline{\boldsymbol{\mu}}$ and $\overline{\boldsymbol{\nu}}$, together with the corresponding directional angle. Results are averaged over five seeds.

Mode	Reference set	$\overline{\boldsymbol{\mu}}^\top \overline{\boldsymbol{\nu}}$	Angle
I2I	COCO-1K	0.590 ± 0.018	53.85°
	Self-val mix-1K	0.482 ± 0.011	61.19°
T2T	COCO captions-1K	0.600 ± 0.009	53.13°
	Medical+Finance-1K	0.475 ± 0.007	61.63°

similarity is 0.948 for $\overline{\boldsymbol{\mu}}$, but only 0.601 for $\overline{\boldsymbol{\nu}}$. The difference is even larger for T2T, where the similarity decreases from 0.874 to 0.445.

These results show that global conformity captures a comparatively stable direction across reference domains, whereas residual conformity is more reference dependent. This effect is particularly pronounced for text representations, where changing from COCO captions to medical and financial text rotates $\overline{\boldsymbol{\nu}}$ by more than 60°.

Table 7: **Cross-set sensitivity of global and residual conformity directions.** For each mode, we compare directions estimated from two different external reference sets using matched seeds.

Mode	Set comparison	Global direction		Residual direction	
		Cosine	Angle	Cosine	Angle
I2I	COCO vs. self-val mix	0.948 ± 0.002	18.58°	0.601 ± 0.016	53.07°
T2T	COCO vs. Medical+Finance	0.874 ± 0.002	29.11°	0.445 ± 0.010	63.56°

A.2 Detailed Implementation of CGD

Having established that residual conformity is distinct from the global conformity direction, we now provide the complete CGD procedure. Algorithm 1 summarizes the complete CGD procedure. CGD first estimates the reference mean, residual conformity direction, and the empirical distribution of residual projection magnitudes from an external reference set. These statistics are fitted once and then kept fixed for all downstream query and gallery embeddings.

Sign of the residual projection. Unlike the conformity projections α_i, which are always positive, the residual projections $\beta_i = \mathbf{z}_i^\top \overline{\boldsymbol{\nu}}$ take both signs: 47.71% of the gallery embeddings for text-to-text (T2T) retrieval and 51.55% for image-to-image (I2I) retrieval have $\beta_i < 0$. So for a query q and a gallery item j, their cosine similarity splits exactly into an orthogonal term and a residual axial term, analogously to the conformity decomposition in Section 3.3. The resultant axial term $\beta_q \beta_j$ increases the similarity of pairs on the same side of the residual axis and decreases it for pairs on opposite sides, independently of their relative orientation in the orthogonal subspace. Its magnitude is $|\beta_q \beta_j|$, which motivates using $|\beta_i|$ to measure the residual strength of an individual embedding.

Empirical percentile. The adaptive correction strength is determined by the position of the current residual magnitude $|\beta_i|$ within the reference distribution. Specifically,

$$p_i = F_{\text{ref}}(|\beta_i|) = \frac{1}{M} \sum_{k=1}^{M} \mathbb{I}\left[|\beta_k^{\text{ref}}| \leq |\beta_i|\right]. \tag{15}$$

Thus, p_i is the fraction of reference embeddings whose residual projection magnitude is no larger than that of the current sample. For example, $p_i = 0.8$ indicates that $|\beta_i|$ is larger than approximately 80% of the residual magnitudes observed in the reference set.

In implementation, the values $\{|\beta_i^{\text{ref}}|\}_{i=1}^M$ are sorted once during fitting. For a downstream embedding, p_i is obtained from the right insertion position of $|\beta_i|$ in this sorted array, divided by M. This

Algorithm 1 Conformity-Guided Debiasing (CGD). All statistics are estimated once on an external reference set and kept frozen at inference; no query or gallery statistics are used.

Input: frozen encoder f of a single modality (image or text);
reference set $\mathcal{B} = \{\mathbf{x}_i^{\text{ref}}\}_{i=1}^{M}$ of the same modality
(default: COCO images for I2I, COCO captions for T2T);
query set $\mathcal{Q} = \{\mathbf{x}_q\}$ and gallery $\mathcal{G} = \{\mathbf{x}_j\}$ (never used for fitting).
Output: corrected embeddings $\mathbf{z}'_q, \mathbf{z}'_j \in \mathbb{S}^{d-1}$ for all $q \in \mathcal{Q}$, $j \in \mathcal{G}$;
a gallery ranking π_q for every query $q \in \mathcal{Q}$.

Stage 1: Fit reference statistics *(offline, once per encoder)*

$\mathbf{u}_i^{\text{ref}} \leftarrow \text{Norm}\big(f(\mathbf{x}_i^{\text{ref}})\big),\ i = 1, \dots, M$ ▷ L_2-normalize
$\boldsymbol{\mu} \leftarrow \frac{1}{M}\sum_{i=1}^{M} \mathbf{u}_i^{\text{ref}}$ ▷ reference mean
$\mathbf{z}_i^{\text{ref}} \leftarrow \text{Norm}(\mathbf{u}_i^{\text{ref}} - \boldsymbol{\mu}),\ i = 1, \dots, M$ ▷ mean correction
$\boldsymbol{\nu} \leftarrow \frac{1}{M}\sum_{i=1}^{M} \mathbf{z}_i^{\text{ref}}, \quad \overline{\boldsymbol{\nu}} \leftarrow \text{Norm}(\boldsymbol{\nu})$ ▷ Eq. (7)
$\beta_i^{\text{ref}} \leftarrow (\mathbf{z}_i^{\text{ref}})^{\top}\overline{\boldsymbol{\nu}},\ i = 1, \dots, M$ ▷ residual projections
$F_{\text{ref}}(a) \leftarrow \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}\big[\,|\beta_i^{\text{ref}}| \le a\,\big]$ ▷ empirical CDF
store and freeze $\Theta = \big(\boldsymbol{\mu}, \overline{\boldsymbol{\nu}}, F_{\text{ref}}\big)$

Stage 2: Correct a single embedding with frozen Θ

function $\text{CGD}(\mathbf{x};\, \Theta)$
  $\mathbf{u} \leftarrow \text{Norm}\big(f(\mathbf{x})\big)$
  $\mathbf{z} \leftarrow \text{Norm}(\mathbf{u} - \boldsymbol{\mu})$ ▷ mean correction, Eq. (5)
  $\beta \leftarrow \mathbf{z}^{\top}\overline{\boldsymbol{\nu}}, \quad \theta \leftarrow \arcsin\beta$ ▷ signed residual angle
  $\mathbf{s} \leftarrow (\mathbf{z} - \beta\,\overline{\boldsymbol{\nu}})\,/\sqrt{1-\beta^2}$ ▷ orthogonal dir., Eq. (9)
  $p \leftarrow F_{\text{ref}}(|\beta|), \quad t \leftarrow (1-p)/2$ ▷ adaptive strength, Eq. (10)
  $\theta' \leftarrow t\,\theta$ ▷ angular contraction
  $\mathbf{z}' \leftarrow \cos\theta'\,\mathbf{s} + \sin\theta'\,\overline{\boldsymbol{\nu}}$ ▷ $= \text{SLERP}(\mathbf{s}, \mathbf{z}; t)$, Eq. (11)
  return $\mathbf{z}'$
end function

Stage 3: Intra-modal retrieval

$\mathbf{z}'_q \leftarrow \text{CGD}(\mathbf{x}_q; \Theta)$ for all $q \in \mathcal{Q}$
$\mathbf{z}'_j \leftarrow \text{CGD}(\mathbf{x}_j; \Theta)$ for all $j \in \mathcal{G}$
for each query $q \in \mathcal{Q}$ **do**
  $s'(q, j) \leftarrow \mathbf{z}'^{\top}_q \mathbf{z}'_j$ for all $j \in \mathcal{G}$ ▷ cosine similarity
  $\pi_q \leftarrow$ gallery indices sorted by $s'(q, \cdot)$ in descending order
end for
return $\{\mathbf{z}'_q\}_{q\in\mathcal{Q}},\ \{\mathbf{z}'_j\}_{j\in\mathcal{G}},\ \{\pi_q\}_{q\in\mathcal{Q}}$

corresponds to the right-continuous empirical CDF in Eq. (15) and allows percentile evaluation by binary search rather than explicitly comparing against all reference samples.

Using $|\beta_i|$ makes the correction symmetric with respect to the two sides of the residual axis: samples with projections $+\beta$ and $-\beta$ of equal magnitude receive the same attenuation strength, while the sign of β_i is preserved through θ_i when reconstructing $\mathbf{z}'_i$. The percentile is converted to the interpolation coefficient $t_i = (1 - p_i)/2$, so samples with larger residual magnitudes receive stronger angular contraction.

Implementation note. For numerical stability, the residual projection is clipped slightly inside $[-1, 1]$ before evaluating both $\arcsin(\beta_i)$ and $\sqrt{1-\beta_i^2}$. The final corrected representation is re-normalized to compensate for floating-point error.

A.3 Effect on Pairwise Similarity

Having detailed the CGD transformation, we next examine how it changes the pairwise cosine similarities used for retrieval.

Analysis 3: How does CGD reshape pairwise similarity?

For the original normalized representations, retrieval uses $s(q, j) = \mathbf{u}_q^\top \mathbf{u}_j$, whereas after CGD the corresponding similarity is $s'(q, j) = \mathbf{z}_q'^\top \mathbf{z}_j'$. Table 8 reports the range and mean of query–gallery similarities on representative I2I and T2T datasets.

CGD consistently shifts the similarity distributions downward. For I2I, the mean similarity decreases from 0.658 to 0.313 on DTD dataset, from 0.643 to 0.405 on Food101, and from 0.816 to 0.637 on EuroSAT. The shift is particularly pronounced for T2T, where the mean similarity decreases from 0.512, 0.457, and 0.487 to 0.000, 0.056, and 0.014 on COCO, Flickr30K, and NoCaps, respectively. The lower minima after CGD further show that pairs are no longer concentrated around a broadly positive similarity offset.

CGD substantially reduces the shared positive similarity level while preserving a broad range of pairwise similarities, consistent with its goal of suppressing conformity-related similarity rather than uniformly rescaling all pairwise scores.

Table 8: **Pairwise cosine similarity before and after CGD.** We report the range and mean of query-gallery similarities together.

Task	Raw range / mean	After CGD range / mean
DTD (I2I)	$[0.262, 1.000]$ / 0.658	$[-0.295, 1.000]$ / 0.313
Food101 (I2I)	$[0.070, 1.000]$ / 0.643	$[-0.254, 1.000]$ / 0.405
EuroSAT (I2I)	$[0.396, 1.000]$ / 0.816	$[0.025, 0.999]$ / 0.637
COCO (T2T)	$[-0.091, 1.000]$ / 0.512	$[-0.438, 1.000]$ / 0.000
Flickr30K (T2T)	$[-0.057, 1.000]$ / 0.457	$[-0.358, 1.000]$ / 0.056
NoCaps (T2T)	$[-0.079, 1.000]$ / 0.487	$[-0.410, 1.000]$ / 0.014

A.4 Visualization of CGD in Real Embedding Spaces

We visualize how the two stages of CGD transform real embedding spaces across datasets and input representations.

Analysis 4: How does CGD modify conformity-related projections in real embedding spaces?

Figure 4 visualizes the two-stage effect of CGD on real embeddings and examines whether the same geometric behavior persists when CGD is applied on top of existing intra-modal corrections. For I2I retrieval, we consider Raw CLIP, IsoCLIP, and PCA$\leftarrow$ representations on Oxford Pets, DTD, and UCF101. For T2T retrieval, we consider Raw CLIP and IsoCLIP representations on COCO Text, Flickr30K Text, and NoCaps Text. PCA$\leftarrow$ is omitted from T2T because its original formulation does not define a text-to-text transformation.

Global conformity before CGD. The left column visualizes the input representation before CGD. Its horizontal coordinate is

$$\alpha = \mathbf{u}^\top \overline{\boldsymbol{\mu}}, \tag{16}$$

i.e., the projection onto the global conformity direction. A non-zero average projection is observed not only for Raw CLIP, but also for representations already transformed by IsoCLIP or PCA$\leftarrow$. This shows that conformity-related directional structure remains present after these existing intra-modal corrections, providing room for an additional conformity-specific correction.

Residual conformity after mean correction. The middle column shows the representation after the first CGD stage,

$$\mathbf{z} = \mathrm{Norm}(\mathbf{u} - \boldsymbol{\mu}). \tag{17}$$

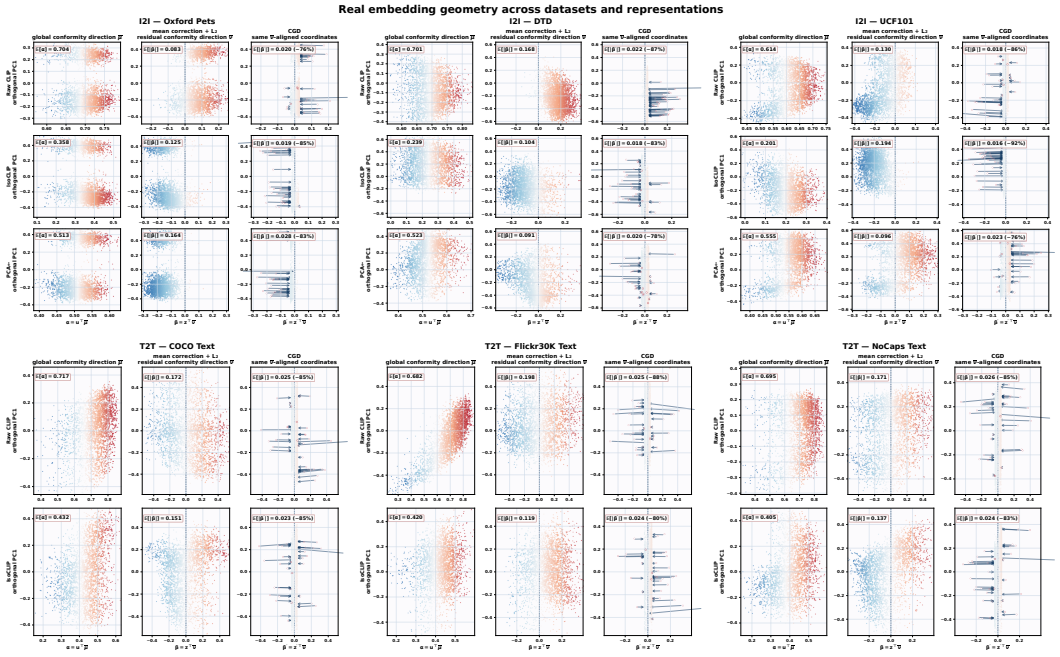


Figure 4: **Geometric effect of CGD on real embedding spaces.** For each representation, the left panel shows the embeddings before CGD, where the horizontal coordinate is the global conformity projection $\alpha = \mathbf{u}^\top \overline{\boldsymbol{\mu}}$. The middle panel shows the representation after mean correction and L_2 re-normalization, where $\beta = \mathbf{z}^\top \overline{\boldsymbol{\nu}}$ is the signed residual conformity projection. The right panel shows the final CGD output in the same $\overline{\boldsymbol{\nu}}$-aligned coordinate system, with $\beta' = \mathbf{z}'^\top \overline{\boldsymbol{\nu}}$. Red circles mark selected samples before the second CGD stage, and arrow endpoints indicate their positions after residual correction; horizontal movement toward zero therefore corresponds directly to attenuation of the residual projection. The vertical coordinate is the leading principal component orthogonal to the corresponding conformity direction and is used only for visualization. Top: I2I retrieval on Oxford Pets, DTD, and UCF101 using Raw CLIP, IsoCLIP, and PCA← representations. Bottom: T2T retrieval on COCO Text, Flickr30K Text, and NoCaps Text using Raw CLIP and IsoCLIP representations. All conformity statistics are estimated from external, representation- and modality-matched COCO reference sets (COCO images for I2I and COCO captions for T2T), without using target query/gallery statistics.

The horizontal coordinate is now the signed residual projection

$$\beta = \mathbf{z}^\top \overline{\boldsymbol{\nu}}. \tag{18}$$

Although the external reference mean has been subtracted, projecting the corrected embeddings back onto the unit sphere through L_2 re-normalization leaves a non-negligible residual projection along $\overline{\boldsymbol{\nu}}$. This behavior is consistently observed across the datasets and input representations shown in the figure.

Residual contraction by the second CGD stage. The right column visualizes the output of the second CGD stage. Importantly, the middle and right panels use exactly the same two-dimensional coordinate system: the horizontal axis is aligned with $\overline{\boldsymbol{\nu}}$, while the vertical axis is the same orthogonal principal component. For each selected sample, the red circle marks its location before residual correction and the arrow endpoint marks its location afterward. Consequently, horizontal movement toward zero directly represents a reduction of the residual conformity component.

We denote the residual projection after CGD by

$$\beta' = \mathbf{z}'^\top \overline{\boldsymbol{\nu}}. \tag{19}$$

Across the examples in Fig. 4, the second CGD stage reduces the mean absolute residual projection from $\mathbb{E}[|\beta|]$ to $\mathbb{E}[|\beta'|]$ by 76–92%. The contraction is observed for Raw CLIP as well as for representations already corrected by IsoCLIP or PCA←.

The visualization supports the two-stage geometric motivation of CGD: conformity-related structure remains in both raw and previously corrected representations; mean correction followed by re-normalization leaves a measurable residual direction; and the second CGD stage substantially contracts this residual component while preserving the orthogonal embedding structure.

B Experimental Details

B.1 Dataset and Evaluation Details

Text-to-text retrieval. Our T2T evaluation contains nine datasets: COCO Text, Flickr30K Text, NoCaps Text, PASCAL-50S, ESCI, WANDS, CQADupStack-English, SciFact, and QuoraRetrieval.

For COCO Text, Flickr30K Text, and NoCaps Text, we follow the T2T retrieval protocol of Iso-CLIP (Magistri et al., 2026). In each dataset, one caption per image is used as the query and the remaining captions associated with the same image form the relevant gallery items. We report full-gallery mAP.

PASCAL-50S does not provide a standard full-gallery T2T retrieval protocol, so we construct one from its 1,000 images and 50 human captions per image. For each image, five captions are selected deterministically as queries and the remaining 45 are used as gallery items. All gallery captions are merged into a single gallery of 45,000 texts, resulting in 5,000 queries, each with 45 relevant captions from the same image. We report full-gallery AP averaged over all queries.

For ESCI, we use the BLaIR-derived Exact-match subset of the U.S. test data. Each query is associated with one labeled target product and is evaluated against 1,367,729 product titles. Relevance is binary in our setup, and we report nDCG@100.

For WANDS, we retrieve over all 42,994 product names. We map *Irrelevant*, *Partial*, and *Exact* to relevance grades 0, 1, and 2, respectively. When multiple annotations exist for the same query–product pair, their numeric grades are averaged. Queries without any positive item are discarded, leaving 479 queries. Unjudged products are assigned relevance zero. We report graded nDCG@10.

For CQADupStack-English, SciFact, and QuoraRetrieval, we use the official test qrels and the complete corpus as gallery. Only queries with at least one positive test qrel are evaluated. CQADupStack-English contains 1,570 evaluated queries and 40,221 corpus questions; SciFact contains 300 claims and 5,183 documents; and QuoraRetrieval contains 10,000 evaluated queries and 522,931 corpus questions. All three use binary relevance and are evaluated with nDCG@10. For CQADupStack-English and SciFact, the corpus representation concatenates the title and text fields.

Image-to-image retrieval. For I2I retrieval, we follow the evaluation protocol used by Iso-CLIP (Magistri et al., 2026) and evaluate on 13 datasets: ROxford5k (Radenovic et al., 2018), RParis6k (Radenovic et al., 2018), CUB (Wah et al., 2025), Stanford Cars (Krause et al., 2013), Oxford-IIIT Pets (Parkhi et al., 2012), Oxford 102 Flowers (Nilsback & Zisserman, 2008), FGVC Aircraft (Maji et al., 2013), SUN397 (Xiao et al., 2010), Caltech101 (Fei-Fei et al., 2007), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), Food101 (Bossard et al., 2014), and UCF101 (Soomro et al., 2012). For ROxford and RParis, we include the R1M distractor set, containing approximately one million negative images, and report the Easy setting. For CUB, the entire dataset is used as both query and gallery. For the remaining datasets, the test split is used as query and the training split as gallery. Relevance is defined by class identity, and performance is measured using mean average precision (mAP).

Image-only few-shot classification. We follow the image-only few-shot protocol of PCA$\leftarrow$ (Herzog & Wang, 2026) on 11 datasets: ImageNet, Caltech101, Oxford-IIIT Pets, Stanford Cars, Oxford 102 Flowers, Food101, FGVC Aircraft, SUN397, DTD, EuroSAT, and UCF101. No text prompts are used. We evaluate 1-, 2-, 4-, 8-, and 16-shot settings with the Prototype and LDA classifiers described in the main paper, and report classification accuracy averaged over the 11 datasets. Results for CGD variants are averaged over three random seeds.

B.2 Alternative Reference Sets

The main experiments use COCO training images and captions as the default reference sets for I2I and T2T, respectively. To examine whether the effectiveness of CGD depends on this particular choice, we replace only the reference set while keeping the downstream evaluation protocol and all other settings unchanged. Specifically, we use ImageNet images as an alternative I2I reference set and Flickr30K training captions as an alternative T2T reference set. We evaluate CLIP ViT-B/32 and SigLIP2 ViT-B/16 as representative backbones.

Table 9: **Alternative reference sets for image-to-image retrieval.** CGD is fitted using either COCO training images (default) or ImageNet images, while the downstream evaluation remains unchanged. Light-blue cells denote representations corrected by CGD, and green superscripts indicate improvements over the corresponding Raw representation. Results are reported in mAP (%).

ViT	Method / Reference	Cal.	CUB	ROxf.	RPar.	Cars	Pets	Flw.	Air.	DTD	Euro.	Food	SUN	UCF	Avg.
CLIP B/32	Raw	77.1	22.9	42.6	67.9	24.6	30.5	62.0	14.5	28.1	47.9	32.3	34.3	47.1	40.9
	+ **CGD** (COCO)	78.8 (+1.7)	23.5 (+0.6)	44.7 (+2.1)	69.7 (+1.8)	25.6 (+1.0)	33.2 (+2.7)	63.6 (+1.6)	14.6 (+0.1)	29.5 (+1.4)	50.4 (+2.5)	34.4 (+2.1)	34.9 (+0.6)	47.3 (+0.2)	42.3 (+1.4)
	+ **CGD** (ImageNet)	79.4 (+2.3)	23.0 (+0.1)	43.9 (+1.3)	69.4 (+1.5)	25.5 (+0.9)	33.9 (+3.4)	63.6 (+1.6)	14.7 (+0.2)	29.5 (+1.4)	50.0 (+2.1)	34.3 (+2.0)	36.0 (+1.7)	47.2 (+0.1)	42.3 (+1.4)
SigLIP2 B/16	Raw	89.2	38.1	53.2	76.5	70.8	56.6	89.3	41.8	39.0	50.8	59.2	43.6	59.2	59.0
	+ **CGD** (COCO)	90.9 (+1.7)	39.4 (+1.3)	54.3 (+1.1)	77.7 (+1.2)	73.1 (+2.3)	59.7 (+3.1)	90.4 (+1.1)	42.4 (+0.6)	39.9 (+0.9)	50.9 (+0.1)	60.1 (+0.9)	44.1 (+0.5)	60.4 (+1.2)	60.3 (+1.2)
	+ **CGD** (ImageNet)	91.4 (+2.2)	39.1 (+1.0)	54.4 (+1.2)	77.5 (+1.0)	72.9 (+2.1)	61.9 (+5.3)	90.5 (+1.2)	41.9 (+0.1)	39.8 (+0.8)	$50.6^{-0.2}$	60.7 (+1.5)	45.3 (+1.7)	59.9 (+0.7)	60.5 (+1.4)

Table 10: **Alternative reference sets for text-to-text retrieval.** CGD is fitted using either COCO training captions (default) or Flickr30K training captions, while the downstream evaluation remains unchanged. Light-blue cells denote representations corrected by CGD, and green superscripts indicate improvements over the corresponding Raw representation. We report the task-native primary metric for each benchmark.

ViT	Method / Reference	COCO mAP (%)	Flickr30K mAP (%)	NoCaps mAP (%)	Pascal-50S mAP (%)	CQADupStack nDCG@10 (%)	ESCI nDCG@100 (%)	Quora nDCG@10 (%)	SciFact nDCG@10 (%)	WANDS graded nDCG@10 (%)
CLIP B/32	Raw	26.2	51.7	35.1	33.6	11.5	6.1	76.2	21.9	54.4
	+ **CGD** (COCO)	29.2 (+3.0)	56.1 (+4.4)	39.6 (+4.5)	38.5 (+4.9)	19.5 (+8.0)	13.3 (+7.2)	77.0 (+0.8)	31.5 (+9.6)	62.1 (+7.7)
	+ **CGD** (Flickr30K)	28.7 (+2.5)	58.2 (+6.5)	39.2 (+4.1)	36.7 (+3.1)	17.6 (+6.1)	11.9 (+5.8)	76.8 (+0.6)	30.5 (+8.6)	60.4 (+6.0)
SigLIP2 B/16	Raw	29.2	56.4	38.8	36.1	12.4	12.4	79.5	24.7	56.8
	+ **CGD** (COCO)	33.4 (+4.2)	62.1 (+5.7)	44.9 (+6.1)	42.8 (+6.7)	16.4 (+4.0)	17.7 (+5.3)	79.9 (+0.4)	30.5 (+5.8)	62.0 (+5.2)
	+ **CGD** (Flickr30K)	31.9 (+2.7)	62.6 (+6.2)	43.2 (+4.4)	39.7 (+3.6)	14.5 (+2.1)	13.6 (+1.2)	79.9 (+0.4)	27.7 (+3.0)	62.5 (+5.7)

Robustness to the reference-set source. Tables 9 and 10 show that the gains of CGD are not tied to the particular COCO reference sets used in the main experiments. For I2I retrieval, replacing COCO with ImageNet yields an average mAP of 42.3 for CLIP ViT-B/32 and 60.5 for SigLIP2 ViT-B/16, compared with 40.9 and 59.0 for their respective Raw representations. These improvements are comparable to those obtained using the default COCO reference set (42.3 and 60.3, respectively). Moreover, with ImageNet as the reference set, CGD improves all 13 I2I benchmarks for CLIP and 12 out of 13 benchmarks for SigLIP2; the only exception is a small 0.2-point decrease on EuroSAT.

The same trend holds for T2T retrieval. When the COCO caption reference set is replaced by Flickr30K training captions, CGD continues to improve the Raw representation on all nine benchmarks for both CLIP and SigLIP2. Although the magnitude of the gain varies across datasets and reference sets, the improvement pattern remains highly stable across both modalities and backbone families.

Overall, these results indicate that CGD does not rely on statistics specific to COCO. Rather, the results are consistent with the reference set providing an estimate of a transferable global geometric bias in the representation space: changing the data used to estimate this bias modifies the exact correction slightly, but preserves its downstream benefit.

B.3 Reference Set Source Construction

To study the effect of the reference-set source, we keep the downstream query/gallery splits and all CGD settings fixed, and vary only the data used to fit CGD. Unless otherwise specified, each reference set contains 1000 samples, drawn without replacement using fixed random seeds. All embeddings are L_2-normalized before fitting CGD. We compare the *default* COCO reference set used in the main experiments with three alternative settings: *in-distribution*, *mixed-distribution*, and *out-of-distribution*.

I2I reference sets. For image-to-image retrieval, we consider 12 datasets: Caltech101, Stanford Cars, Oxford-IIIT Pets, Oxford Flowers, FGVC Aircraft, DTD, EuroSAT, Food101, UCF101, Country211, STL10, and SUN397. For each dataset, we construct a reference candidate pool that is disjoint from the query/gallery images used for retrieval evaluation. Most datasets use their official validation split; Country211 uses the official valid split, STL10 uses the unlabeled split, and SUN397 uses images that are not included in the query/gallery split.

For a given I2I retrieval benchmark, the four settings are defined as follows:

1) Default: samples are drawn from the COCO training images.

2) In-distribution: samples are drawn from the held-out pool of the same retrieval benchmark. The reference data therefore follow the same distribution as the downstream task, while remaining disjoint from the evaluation query/gallery set.

3) Mixed-distribution: the reference set is constructed by balancing samples across the held-out pools of all 12 I2I datasets, including the current retrieval benchmark.

4) Out-of-distribution: the held-out pool of the current retrieval benchmark is excluded, and the set is constructed from the remaining 11 image datasets.

T2T reference sets. For text-to-text retrieval, we conduct the same source ablation on COCO Text, Flickr30K Text, and NoCaps Text. Each benchmark provides a reference text pool that is separated from the retrieval evaluation data: COCO uses val_query, Flickr30K uses train_query, and NoCaps uses a fixed set of 1000 disjoint queries.

The four T2T settings are:

1)Default: samples are drawn from the first captions of the COCO training set.

2)In-distribution: samples are drawn from the independent reference pool of the same text retrieval benchmark.

3)Mixed-distribution: the reference set is formed by approximately balancing samples from the COCO, Flickr30K, and NoCaps reference pools.

4)Out-of-distribution: the set contains 500 medical queries from NFCorpus and 500 financial queries from FiQA, providing a deliberately cross-domain text reference.

C Additional Ablations

We further examine three design choices of CGD: whether residual conformity should be corrected on both sides of the residual direction, how performance changes with the amount of residual angle retained, and whether repeatedly refitting CGD provides additional benefits.

Residual correction design. Recall that the signed residual projection is

$$\beta_i = \mathbf{z}_i^\top \overline{\boldsymbol{\nu}}. \tag{20}$$

The proposed CGD correction is bidirectional: the correction strength is determined by $|\beta_i|$, while the sign of β_i is preserved through the signed angle $\theta_i = \arcsin(\beta_i)$.

To examine whether both sides of the residual direction should be corrected, we compare three variants. The *positive-only* variant applies the second-stage angular correction only to samples with $\beta_i > 0$ and leaves samples with $\beta_i \leq 0$ unchanged. The *negative-only* variant analogously corrects only samples with $\beta_i < 0$. The proposed *bidirectional* variant applies the same adaptive correction to samples on both sides of $\overline{\boldsymbol{\nu}}$.

We evaluate CLIP ViT-B/32 on three I2I datasets (EuroSAT, Oxford Pets, and Stanford Cars) and three T2T datasets (COCO, Flickr30K, and NoCaps). As shown in Fig. 5(a), either one-sided variant can improve retrieval, but bidirectional correction gives the largest gain on all six datasets. The difference is particularly pronounced on Flickr30K, where the positive-only and negative-only variants improve mAP by $+0.9$ and $+2.5$, respectively, compared with $+4.4$ for bidirectional correction.

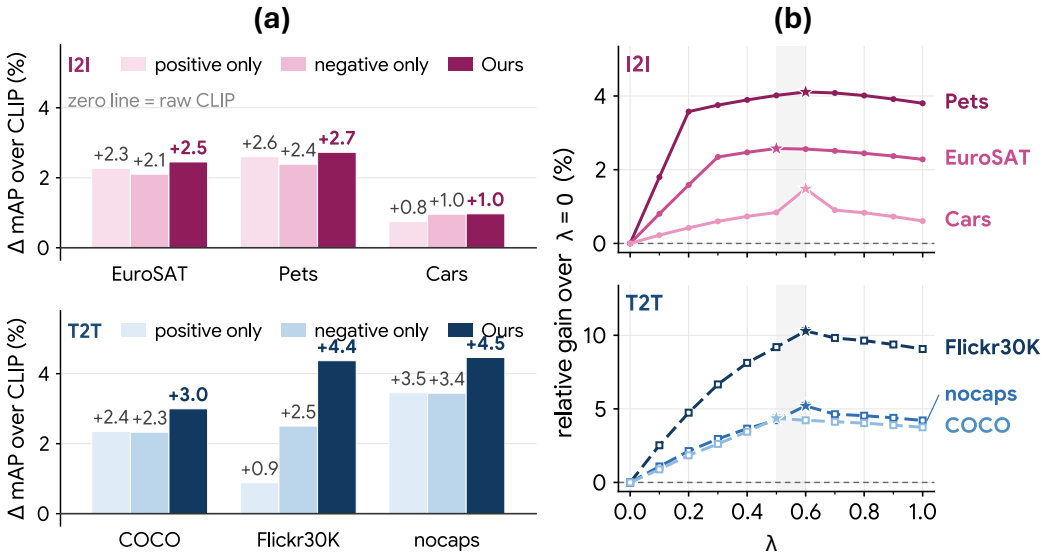


Figure 5: **Additional ablations of the residual conformity correction** (CLIP ViT-B/32). **(a)** One-sided vs. bidirectional correction on three I2I and three T2T benchmarks. The positive-only (negative-only) variant applies the second-stage angular correction only to samples with $\beta_i > 0$ ($\beta_i < 0$) and leaves the remaining samples unchanged, whereas CGD (Ours) corrects both sides of $\bar{\boldsymbol{\nu}}$. Bars report the absolute mAP gain (in points) over raw CLIP. **(b)** Sensitivity to the angular-retention coefficient λ in $t_i^{(\lambda)} = \lambda(1 - p_i)$ (Eq. equation 21); $\lambda = 0$ corresponds to the hard projection and $\lambda = 0.5$ recovers the default CGD. All other settings are fixed. The y-axis reports the relative mAP improvement over $\lambda = 0$, i.e., $(\mathrm{mAP}_\lambda - \mathrm{mAP}_0)/\mathrm{mAP}_0 \times 100\%$. The shaded band and star markers indicate the default $\lambda = 0.5$.

Residual conformity affects retrieval on both sides of the residual direction, supporting the use of $|\beta_i|$ to determine the adaptive correction while preserving the sign of β_i during reconstruction.

Sensitivity to angular retention. The default CGD formulation uses $t_i = (1 - p_i)/2$ to retain a sample-dependent fraction of the signed residual angle. To examine the sensitivity to this retention strength, we introduce a global retention scale $\lambda \in [0, 1]$ and generalize the interpolation coefficient as

$$t_i^{(\lambda)} = \lambda(1 - p_i). \tag{21}$$

The default CGD formulation is recovered at $\lambda = 0.5$. Smaller values of λ induce stronger contraction toward the subspace orthogonal to $\bar{\boldsymbol{\nu}}$; in particular, $\lambda = 0$ reduces to hard projection. Larger values retain a greater fraction of the original residual angle.

Figure 5(b) reports the performance change relative to the hard-projection setting $\lambda = 0$. Across both I2I and T2T datasets, performance improves substantially once part of the residual angle is retained and remains relatively stable over a broad intermediate range. Performance decreases slightly when too much residual information is retained, indicating that neither complete removal nor weak attenuation is optimal.

The results support the partial residual correction used by CGD: hard projection is unnecessarily aggressive, while retaining a moderate fraction of the residual angle consistently improves retrieval. The performance is also stable around the default setting.

Iterative CGD. The default method applies CGD once. Since each CGD round changes the representation geometry, we further examine whether repeatedly re-estimating the conformity statistics in the corrected space provides additional retrieval gains, and how this behavior relates to the remaining conformity of the downstream representations.

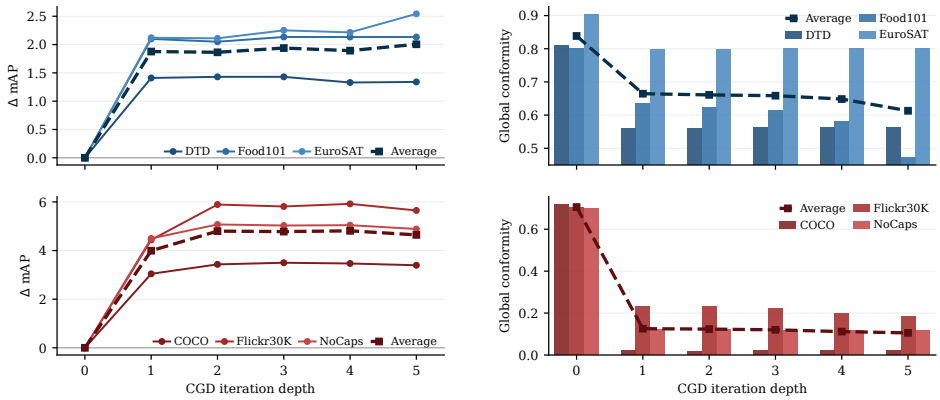


Figure 6: **Iterative CGD and dataset-internal conformity.** Left: retrieval improvement, measured as the mAP gain over depth 0, across successive CGD rounds. Right: dataset-internal conformity $C^{(t)} = \left\| \frac{1}{|\mathcal{X}^{(t)}|} \sum_{\mathbf{x} \in \mathcal{X}^{(t)}} \mathbf{x} \right\|_2$ measured on the normalized union of downstream query and gallery embeddings. The top row reports I2I results on DTD, Food101, and EuroSAT, and the bottom row reports T2T results on COCO, Flickr30K, and NoCaps. Dashed curves denote the average over the three datasets. Depth 0 corresponds to the original CLIP representation; at each subsequent depth, CGD is refitted on the transformed external COCO reference set before being applied again. The downstream conformity measure is used only for post-hoc analysis and is never used to fit CGD.

At iteration t, CGD is fitted only on the current external COCO reference set $\mathcal{B}^{(t)}$. After applying the fitted transformation to the reference embeddings, we construct

$$\mathcal{B}^{(t+1)} = \mathrm{CGD}^{(t)}\left(\mathcal{B}^{(t)}\right), \tag{22}$$

and re-estimate the reference mean $\boldsymbol{\mu}^{(t+1)}$, residual conformity direction $\overline{\boldsymbol{\nu}}^{(t+1)}$, and empirical residual-magnitude distribution $F_{\mathrm{ref}}^{(t+1)}$ for the next round. The same sequence of fitted transformations is applied to the downstream embeddings, but no downstream query or gallery statistics are used for fitting.

We evaluate depths 0–5 on three representative I2I datasets (DTD, Food101, and EuroSAT) and three T2T datasets (COCO, Flickr30K, and NoCaps). Depth 0 denotes the original CLIP representation, while depth t corresponds to applying t successively refitted CGD rounds.

In addition to retrieval performance, we measure the remaining *dataset-internal global conformity* after each iteration. For the normalized union of downstream query and gallery embeddings, $\mathcal{X}^{(t)} = \mathcal{Q}^{(t)} \cup \mathcal{G}^{(t)}$, we compute

$$C^{(t)} = \left\| \frac{1}{|\mathcal{X}^{(t)}|} \sum_{\mathbf{x} \in \mathcal{X}^{(t)}} \mathbf{x} \right\|_2 . \tag{23}$$

Since the embeddings are L_2-normalized, $C^{(t)}$ is equivalently the average projection onto the normalized mean direction of the downstream embedding set. Importantly, this quantity is used only as a post-hoc diagnostic and never participates in CGD fitting.

As shown in Fig. 6, the first CGD round produces both a large retrieval improvement and a sharp reduction in dataset-internal global conformity. This effect is particularly pronounced for T2T, where conformity drops substantially after the first iteration while most of the retrieval gain is already obtained. Additional rounds further reduce conformity, but the corresponding retrieval improvements become considerably smaller and are not monotonic across datasets. For I2I, the retrieval gain increases more gradually, whereas global conformity continues to decrease across iterations.

Most of the retrieval benefit is obtained in the first one or two CGD rounds, whereas global conformity can continue to decrease afterward. We therefore use a single CGD round throughout the main experiments, preserving the simple one-shot, plug-and-play formulation without requiring an iteration-specific stopping rule.