GAPDIM: DIMENSION PRUNING FOR MODALITY GAP REDUCTION IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive Vision-Language Models (VLMs) such as CLIP and SigLIP exhibit a modality gap, whereby embeddings from different modalities occupy separate regions of the shared latent space. Although this fragmentation does not necessarily prevent strong retrieval performance, it limits the use of pretrained embeddings for out-of-distribution or group-wise tasks such as clustering. Existing methods to reduce the modality gap typically require retraining with additional losses or architectural changes, while post-hoc alternatives often degrade retrieval by disrupting the instance-wise alignment learned during pretraining. In this paper, we ask whether the portion of the gap that disrupts the multimodal organization of the representation space can be removed selectively. We observe that the gap is concentrated in a few coordinates of the native bases of pretrained models and that these coordinates exhibit low participation in the dominant centered subspaces. Based on this analysis, we propose Gap-Importance DIMension pruning (GapDIM), a post-hoc, training-free method that identifies coordinates with high modality gap and low subspace participation and selectively prunes them from the embedding space. GapDIM thereby reduces the modality gap while preserving the discriminative structure of pretrained representations. Across multiple VLM backbones and datasets, GapDIM generally preserves retrieval performance while reducing modality separation and improving joint clustering in most configurations. These results suggest that the largest part of the modality gap is concentrated in a small set of low-participation embedding coordinates, enabling a training-free calibration strategy. Code available at: `https://anonymous.4open.science/r/GapDIM-3DDB`.

1 INTRODUCTION

Contrastive Vision-language models (VLMs), including CLIP Radford et al. (2021), BLIP Li et al. (2023), SigLIP Zhai et al. (2023); Tschannen et al. (2025), and Flamingo Alayrac et al. (2022), map text and images into a shared representation space. Although their training procedure induces instance-level alignment, their embeddings often remain separated according to the modality rather than being grouped by semantics, a phenomenon known as *modality gap* Liang et al. (2022).

Several works propose methods to mitigate the modality gap by fixing the temperature Yaras et al. (2025) of the InfoNCE loss van den Oord et al. (2018), including additional losses Grassucci et al. (2026); Liu et al. (2026); Fahim et al. (2024); Shi et al. (2023), architectural modifications Eslami & de Melo (2025), or by diffusion models Lee et al. (2025). Nevertheless, the impact of these methods in retrieval tasks is mostly irrelevant or still under debate Chowers et al. (2026). However, such strategies require the full training set and large amounts of computational resources, which are often unfeasible for very large or proprietary models and datasets. This limits the applicability of loss- or architecture-based methods to reduce the modality gap. To address the applicability limitation, recently, some works have proposed training-free methods to close the gap. These include rigid translations of the embeddings Liang et al. (2022) or through procrustes D'Orazio et al. (2026), matching statistics and subspace alignment Yu et al. (2026), global normalization of embeddings An et al. (2025) and removal of modality-specific information Li et al. (2025). Although computationally light and easy to apply to pretrained models, these methods close the gap without taking care of the instance-wise alignment achieved during training. As a consequence, while clustering im-

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.

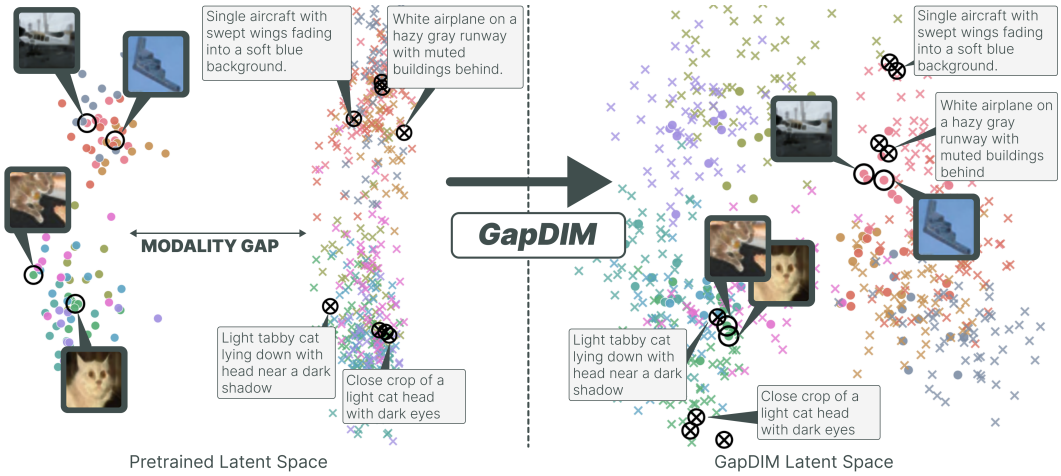

Figure 1: Latent space of the pretrained ViT-bigG/14 model (left) and after applying GapDIM (right). GapDIM reduces the modality gap, revealing multimodal semantic clusters.

proves, retrieval performance often severely degrades, undermining their effective usage in practical scenarios.

Our work addresses these limitations directly. Building on evidence that the gap can concentrate in a small number of embedding coordinates Schrodi et al. (2025), we ask whether these coordinates also participate strongly in the dominant variation of the two modalities. We find that coordinates with a high gap in the learned latent space tend to exhibit low participation in the modalities-shared dominant subspace, and are empirically associated with removable gap noise. Therefore, we argue that these coordinates can be filtered out without compromising VLMs performance. This analysis motivates Gap-Importance DIMension pruning (GapDIM), a novel training-free method that identifies coordinates with high modality discrepancy and low dominant-subspace participation through a gap-leverage bound that formalizes their coordinate-wise relationship. This allows GapDIM to prune high-gap, low-participation coordinates in the pretrained model native basis. This targeted pruning is what enables GapDIM to address prior training-free method limitations: reducing the modality gap while matching or, more often, improving the retrieval performance of the pretrained VLM, alongside a considerable improvement in clustering (up to +9 points in V-measure). GapDIM results are clearly visible in Fig. 1, in which the pretrained space of the VLM is fragmented, exhibiting modality gap, while after applying the proposed GapDIM the gap is reduced and multimodal semantic clusters appear more homogeneously. The entire GapDIM pipeline is post-hoc and training-free, requires no update of the encoder parameters, and is directly applicable to any pretrained VLM, including CLIP and SigLIP variants of different sizes. In our extensive experimental evaluation, we show that our findings are consistent across multiple datasets and VLMs of different parameter numbers and embedding sizes.

The main contributions of this work can be summarized as follows:

- We provide a thorough formulation of the emergence of high-gap, low-participation coordinates in contrastive VLM representations and derive a bound showing that the fraction of the modality gap carried by an embedding coordinate is controlled by how strongly that coordinate participates in each modality dominant centered subspace.
- We propose Gap-Importance DIMension pruning (GapDIM), a post-hoc method that identifies coordinates with high gap and low dominant-subspace participation. GapDIM selectively prunes these coordinates without updating or fine-tuning the pretrained encoders.
- Throughout our evaluation, we show that the presence of high-gap and low-importance dimensions is common to several contrastive VLMs variants from small- to large-scale and with different embedding sizes, and that GapDIM is effective in each of these scenarios.

2 Related Work

Contrastive Vision-Language Models (VLMs). Contrastive VLMs such as CLIP Radford et al. (2021) and SigLIP Zhai et al. (2023); Tschannen et al. (2025) typically use one encoder per modality, jointly trained on large corpora of image-text pairs. CLIP employs the InfoNCE loss van den Oord et al. (2018), which pushes paired embeddings together while repelling all other pairs in the batch. SigLIP Zhai et al. (2023); Tschannen et al. (2025) replaces the InfoNCE softmax-based normalization with a sigmoid loss that operates on image-text pairs independently, making training more suitable for large-scale distributed settings. Nevertheless, standard contrastive objectives do not explicitly require the two modalities to occupy the same global region of the shared latent space Grassucci et al. (2026); Yaras et al. (2025).

Modality gap and training regularization. The modality gap was characterized in Liang et al. (2022): multimodal contrastive models project the two modalities into separate regions of the shared embedding space, yielding a systematic displacement between matching pairs caused by the cone effect. Subsequent work Schrodi et al. (2025) shows that the gap primarily arises from information imbalance between modalities, i.e., the availability of a richer semantic signal in one modality than the other, and further establishes that only a few dimensions of the embedding space are actually responsible for most of the gap. Although several works propose methods to close such a gap Eslami & de Melo (2025); Liu et al. (2026); Grassucci et al. (2026); Shi et al. (2023); Xu et al. (2025); Grassucci et al. (2025); Huang et al. (2025), a natural question is whether the gap should be closed or left as it is Ramasinghe et al. (2024). Closing the gap is useful for group-wise tasks such as clustering, in which absolute placement of embeddings matter Grassucci et al. (2026), for out-of-distribution detection Wang et al. (2026), or for semantic communications Grassucci et al. (2025). However, closing the gap at training time is non-trivial: retraining the encoders from scratch is expensive and requires the availability of large datasets. These observations motivate the development of post-hoc, training-free alternatives.

Modality gap and post-hoc methods. Several recent methods address the modality gap without updating model parameters. Embedding Shift Liang et al. (2022) rigidly translates one modality along the gap vector to align the two centroids. However, it consistently disrupts the fine-grained structure of the embedding space, degrading retrieval. ReAlign Yu et al. (2026) matches first- and second-order statistics between modalities in the normalized space. Although effective at reducing global discrepancies, the enforced statistical alignment alters the embedding structure, impacting retrieval performance. GR-CLIP Li et al. (2025) suppresses modality-specific components to align the two distributions. This post hoc transformation is the same used in I0T An et al. (2025), which standardizes each modality embeddings dimension-wise. However, this global normalization does not distinguish semantic from non-semantic dimensions, so informative structure may be inadvertently altered alongside modality-specific noise.

All of these methods share a common problem: they close the gap without explicitly and perfectly preserving cross-modal alignment, which degrades retrieval performance. In contrast to all the above, our approach selectively identifies and prunes the dimensions that carry both a high gap and low semantic importance, preserving the discriminative structure of the embedding space while reducing fragmentation. As a result, we generally match or surpass pretrained retrieval performance with a minimal computational overhead, and improve clustering in most tested configurations.

Subspace and coordinate selection. Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) are commonly used to identify dominant low-rank structure in learned representations Jolliffe & Cadima (2016); Halko et al. (2011). Truncated SVD represents a data matrix through its leading singular directions and, according to the Eckart-Young-Mirsky theorem, provides the optimal rank-constrained approximation under standard matrix norms. Feature leverage scores bridge a principal subspace to the original coordinates by measuring the diagonal of its orthogonal projector Drineas et al. (2012). They are therefore widely used in feature and column subset selection to transfer information from singular directions back to the original variables.

3 Gap Geometry in Vision-Language Representations

In this section, we introduce the geometric quantities underlying GapDIM and investigate why coordinates carrying a large modality gap can exhibit limited participation in the dominant centered subspaces.

3.1 BACKGROUND

Notation. Throughout the paper, we consider two modalities X and Y, with respective normalized paired embedding matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{N \times D}$, being N the number of samples and D the embedding dimension. Accordingly, $M \in \{X, Y\}$ denotes one of the two modalities. The embeddings are extracted through two encoders f_θ and g_ϕ , respectively, which produce a text embedding $\mathbf{x} = g_\phi(T), \mathbf{x} \in \mathbb{R}^D$ and an image embedding $\mathbf{y} = f_\theta(I), \mathbf{y} \in \mathbb{R}^D$ for a given text T and an image I. The embeddings are then normalized to unitary norm. All coordinate-wise operations are performed in the native embedding basis produced by the pretrained projection heads.

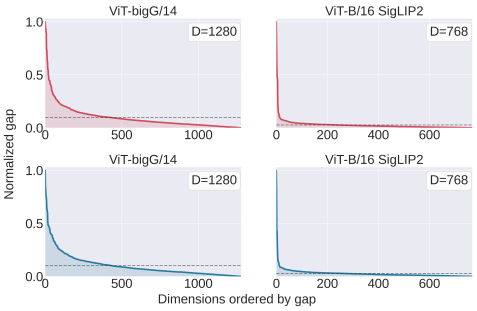

Figure 2: Normalized gap ordered by embedding dimensions. Few dimensions of the modalities-shared latent space from pretrained models drive the modality gap.

Contrastive vision-language models. As reviewed in Section 2, contrastive VLMs optimize the relative compatibility of paired text and image embeddings. Because the embeddings are normalized to norm 1, their cosine similarity $\mathrm{sim}(\cdot, \cdot)$ is simply the dot product. CLIP Radford et al. (2021) uses the symmetric InfoNCE objective:

$$\mathcal{L}_{\mathrm{CLIP}} = \frac{1}{2}\left(\mathcal{L}_{X \to Y} + \mathcal{L}_{Y \to X}\right), \tag{1}$$

where

$$\mathcal{L}_{X \to Y} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\mathrm{sim}(\mathbf{x}_i, \mathbf{y}_i)/\tau)}{\sum_{j \in \mathcal{J}, j \neq i}^{N} \exp(\mathrm{sim}(\mathbf{x}_i, \mathbf{y}_j)/\tau)}, \tag{2}$$

and $\mathcal{L}_{Y \to X}$ is defined symmetrically. SigLIP Zhai et al. (2023); Tschannen et al. (2025) replaces the batch-normalized softmax objective with a pairwise sigmoid loss. These objectives constrain relative similarities, but do not uniquely determine the absolute region occupied by each modality. Strong pairwise retrieval can therefore coexist with globally separated modality distributions Liang et al. (2022); Grassucci et al. (2026); Yaras et al. (2025).

Modality gap. The modality gap can be measured differently, with each metric considering different aspects of the fragmented space, such as centroid distance and overall used space, and the effective similarity between paired image-text samples. L2-distance between the means (L2M) is the original method to measure the modality gap Liang et al. (2022), and it is defined as:

$$\mathrm{L2M} := \|\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y\|, \qquad \boldsymbol{\mu}_X = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i, \qquad \boldsymbol{\mu}_Y = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i, \tag{3}$$

where $\boldsymbol{\mu}_X, \boldsymbol{\mu}_Y \in \mathbb{R}^D$ are respectively the centroids of modality X and Y. We additionally consider the scale-aware Relative Modality Gap (RMG), which is defined in Appendix H.3.1.

3.2 GAPDIM

The proposed Gap-importance DIMension pruning (GapDIM) method builds on the observation that the modality gap is mostly concentrated in a small number of embedding coordinates, and that these dimensions account for most of the overall modality gap magnitude Schrodi et al. (2025). Figure 2 shows the typical scenario in which few dimensions carry most of the gap. GapDIM identifies and prunes such dimensions to build a more coherent latent space. The proposed GapDIM method is composed of three phases, illustrated in Fig. 3 and detailed below.

Phase 1: Gap computation. Given the set of normalized paired embeddings $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{N \times D}$, the first step is to compute the centroids of each modality as equation 3 and then define the per-coordinate gap as

$$g_j = |\boldsymbol{\mu}_{X,j} - \boldsymbol{\mu}_{Y,j}|, \qquad j = 1, \ldots, D. \tag{4}$$

where j represents the indexed coordinate in the embedding space. The resulting vector is normalized to obtain $g_j \in [0, 1]$. With this formulation, we quantify the cross-modal mismatch at the level

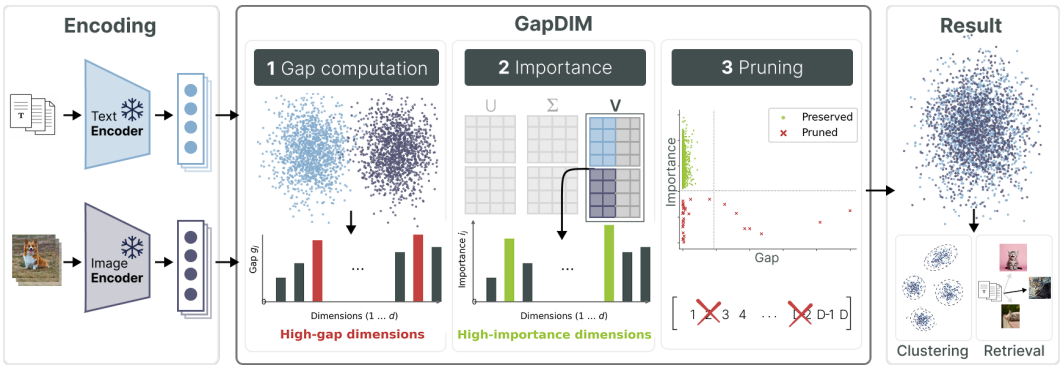

Figure 3: The three-phase GapDIM framework.

of individual embedding coordinates and estimate the relative contribution of each coordinate to the overall modality gap.

Phase 2: Importance computation. The modality gap describes a difference in absolute position, in the second step we study sample-dependent variation independently of this displacement, hence we measure how much each coordinate contributes to the principal subspace. Here D_{sub} denotes the number of principal directions used to estimate coordinate participation, which is different from the number of final coordinates retained after pruning, which is $|\mathcal{I}|$, defined later. To do so, we first center both modalities as:

$$\mathbf{X}_c = \mathbf{X} - \mathbf{1}\boldsymbol{\mu}_X^\top, \qquad \mathbf{Y}_c = \mathbf{Y} - \mathbf{1}\boldsymbol{\mu}_Y^\top \tag{5}$$

where $\mathbf{1} \in \mathbb{R}^N$ is a column vector of ones. Subtracting this term from $\mathbf{X}$ centers the data by removing the global mean from each sample. This operation ensures that the subsequent Singular Value Decomposition (SVD) captures directions of maximal variance rather than being influenced by the absolute position of the data in the feature space. Then, we decompose the centered matrices through SVD:

$$\mathbf{X}_c = \mathbf{U}_X \mathbf{\Sigma}_X \mathbf{V}_X^\top, \qquad \mathbf{Y}_c = \mathbf{U}_Y \mathbf{\Sigma}_Y \mathbf{V}_Y^\top, \tag{6}$$

with $\mathbf{U} \in \mathbb{R}^{N\times N}, \mathbf{\Sigma} \in \mathbb{R}^{N\times D}, \mathbf{V} \in \mathbb{R}^{D\times D}$, in which singular values of $\mathbf{\Sigma}$ quantify directional energy concentrations of the feature directions of $\mathbf{V}$. This choice is motivated by the Eckart-Young-Mirsky theorem and the standard PCA interpretation of the SVD Jolliffe & Cadima (2016); Halko et al. (2011). For a centered matrix $\mathbf{A}_c = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, the rank-$D_{\text{sub}}$ truncation $\mathbf{A}_{c,D_{\text{sub}}} = \mathbf{U}_{D_{\text{sub}}} \mathbf{\Sigma}_{D_{\text{sub}}} \mathbf{V}_{D_{\text{sub}}}^\top$ is the best rank-D_{sub} approximation in spectral norm, with

$$\|\mathbf{A}_c - \mathbf{A}_{c,D_{\text{sub}}}\|_2 = \sigma_{D_{\text{sub}}+1}.$$

where σ_i is the i-th eigenvalue of matrix $\mathbf{A}_c$. Thus, the leading right singular vectors span the dominant low-rank subspace of each modality. In $\mathbf{V}$, the principal directions corresponding to larger singular values capture most of the dominant feature variation, while the ones associated with smaller singular values often encode residual or redundant variation Guo et al. (2026); Li et al. (2026); Magistri et al. (2026). Given this theoretical guarantee, we extract the top-D_{sub} directions encoding strong and middle signal information via the interquartile range of the singular values as a practical implementation, following Hartford et al. (2024)Guo et al. (2026). This results in:

$$\mathbf{W}_M = [\mathbf{v}_{M,1} \quad \cdots \quad \mathbf{v}_{M,D_{\text{sub}}}] \in \mathbb{R}^{D\times D_{\text{sub}}}, \qquad M \in \{X, Y\}, \tag{7}$$

which contains the first D_{sub} right singular vectors of the centered embedding matrix for modality M, whose columns are orthonormal.

Phase 3: Dimension Pruning. Given $\mathbf{W}_M$, we compute the coordinate leverage score Drineas et al. (2012), which measures how strongly an original coordinate participates in the dominant centered subspace of a modality. It is computed as following:

$$\ell_{M,j} = \|\mathbf{W}_M^\top \mathbf{e}_j\|_2^2 = \mathbf{e}_j^\top (\mathbf{W}_M \mathbf{W}_M^\top) \mathbf{e}_j = (\mathbf{W}_M \mathbf{W}_M^\top)_{jj}, \tag{8}$$

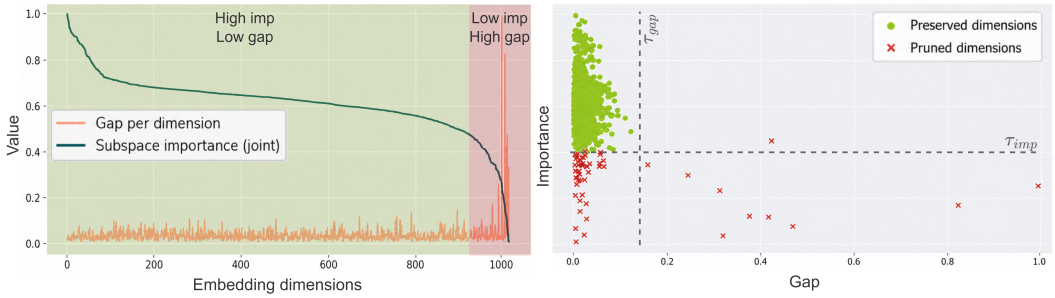

Figure 4: Embedding-coordinate analysis on MSCOCO. Left: coordinates ordered by decreasing participation score; low-participation coordinates exhibit the largest gap values. Right: retained and pruned coordinates according to the thresholds τ_{imp} and τ_{gap}.

where $\mathbf{e}_j$ is the j-th canonical basis vector. These quantities are diagonals of the projection matrix onto the top-D_{sub} right-singular subspace Drineas et al. (2012). The joint importance is then the average over the leverage scores of the two modalities:

$$\ell_j = \frac{1}{2}(\ell_{X,j} + \ell_{Y,j}), \tag{9}$$

which is normalized yielding $\ell_j \in [0, 1]$. Low values indicate low participation in the dominant subspace and are empirically associated with removable gap noise in our experiments.

Once we know the subspace participation and gap score of each dimension j, we retain coordinates that satisfy both thresholds τ_{imp} and τ_{gap}:

$$\mathcal{I} = \{j \in \{1, \ldots, D\} \mid \ell_j \geq \tau_{\text{imp}} \ \wedge \ g_j \leq \tau_{\text{gap}}\}. \tag{10}$$

The selected dimensions are applied to the original embeddings by restricting them to the index set $\mathcal{I}$: $\mathbf{X}^{\text{sel}} = [\mathbf{X}]_{\mathcal{I}}, \mathbf{Y}^{\text{sel}} = [\mathbf{Y}]_{\mathcal{I}}$, yielding the final embeddings with high-gap and low-subspace-participation dimensions pruned.

3.3 Theoretical Analysis

We now validate the theoretical aspects of GapDIM by characterizing the image-text representation space following Zhang et al. (2024), which models the difference between paired representations as $\mathbf{x}_i - \mathbf{y}_i = \mathbf{c}_\perp + \boldsymbol{\epsilon}_i$, where $\mathbf{c}_\perp$ is an approximately constant modality-dependent offset and $\boldsymbol{\epsilon}_i$ represents sample-dependent alignment noise. From this formulation, we yield the following propositions.

Proposition 1 (invariance to modality-dependent translations). Let every embedding of modality M be translated by the same vector $\mathbf{c}_M$, then, the centered embedding matrix in equation 5, and consequently its right singular subspace in equation 7, and its coordinate leverage scores in equation 8 remain unchanged. In contrast, the modality-gap vector changes by $\mathbf{c}_X - \mathbf{c}_Y$. This explains how a coordinate can carry a large modality gap without contributing substantially to the sample-dependent centered variation. A modality gap changes the centroids but it may be exactly canceled by the centering operation used to estimate $\mathbf{W}_M$ in equation 5. Therefore, Proposition 1 provides an intuitive explanation for the decoupling of the gap and importance quantities.

Proposition 2 (coordinate gap-leverage bound). We define coordinate j's gap energy as the fraction of the total squared gap it carries:

$$a_j = \frac{g_j^2}{\|\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y\|_2^2}, \qquad \sum_{j=1}^{D} a_j = 1. \tag{11}$$

Let

$$\mathbf{u} = \frac{\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y}{\|\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y\|_2} \tag{12}$$

denote the normalized gap direction. For each modality, we define the subspace overlap

$$\rho_M = \|\mathbf{P}_M \mathbf{u}\|_2^2 = \frac{\|\mathbf{W}_M^\top (\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y)\|_2^2}{\|\boldsymbol{\mu}_X - \boldsymbol{\mu}_Y\|_2^2}, \qquad \rho_M \in [0, 1]. \tag{13}$$

The value ρ_M is the fraction of the total gap energy contained in the dominant subspace of modality M. The case $\rho_M = 0$ corresponds to exact orthogonality, whereas $\rho_M = 1$ means that the gap lies entirely inside that subspace. Therefore, for every modality M and coordinate j, we have:

$$\sqrt{a_j} \leq \sqrt{\ell_{M,j}\rho_M} + \sqrt{(1 - \ell_{M,j})(1 - \rho_M)}, \tag{14}$$

which means that, for a given coordinate j, its gap energy is upper-bounded by its importance and the fraction of the gap lying inside the selected subspace. Interestingly, in the case of the orthogonal gap as described in Zhang et al. (2024), equation 14 simplifies to $a_j \leq 1 - \ell_{M,j}$. Therefore, the j-th coordinate gap energy is bounded by the inverse of its leverage score. This establishes a relation between the gap and the leverage score of each coordinate. We provide the derivation of equation 14 in Appendix A and also an analysis of the non-orthogonal gap case in Appendix C.

4 Experimental Results

4.1 Experimental Setup

Tasks and backbones. We evaluate the transformed embedding spaces on two downstream tasks that capture complementary properties of multimodal representations: cross-modal retrieval and clustering, and on three datasets of different scales: Flickr30k, MSCOCO, and LLaVA-CC3M, for which the details are provided in Appendix H. Additionally, we perform image-free zero-shot image captioning on MSCOCO. Pipeline details in Appendix I.

Backbones and comparisons. We evaluate multiple vision-language backbones spanning different architectures, parameter counts, and embedding dimensions. Specifically, we consider OpenCLIP models based on ViT-B/32, ViT-B/16, and ViT-bigG/14, together with the SigLIP2 variants ViT-B/16-SigLIP2-256 and ViT-L/16-SigLIP2-256. Their embedding dimensions D are 512, 512, 1280, 768, and 1024, respectively, and their approximate parameter counts are 151M, 150M, 2.5B, 86M, and 303M. We compare GapDIM with post-hoc, training-free methods that require neither encoder retraining nor the training of additional models: Embedding Shift Liang et al. (2022), ReAlign Yu et al. (2026), and I0T An et al. (2025).

Threshold setting. As discussed in section 3.2, for each dimension of the embedding, we estimate the importance that it assumes in the subspace. We perform the optimization of the two thresholds τ_{imp} and τ_{gap} Optuna with a TPESampler for automating threshold tuning. Since our aim is to preserve as much as possible retrieval performance, being the most common task for VLMs, the objective function to be optimized for dimension removal is directly R@1. Details in Appendix D.1. We further devise a generic and unsupervised threshold rule in Appendix D.2.

4.2 Results Discussion

Coordinate distribution. The coordinate-wise gap and participation scores exhibit three recurring properties across the evaluated model-dataset configurations. First, fewer than 3% of the embedding coordinates carry most of the gap. Second, these high-gap coordinates exhibit low participation in the dominant centered subspaces, as measured by the joint score ℓ_j. Third, no coordinate simultaneously exhibits high gap and high participation, as visible in Fig. 4, no points appear in the corresponding upper-right quadrant. This last observation is very significant: it implies that the gap and the semantic structure of the representations mostly live in subspaces identified by different embedding coordinates, consistent with Zhang et al. (2024), and that the modality-gap can therefore be mitigated selectively without touching discriminative content. This is consistent with the fact that, despite not closing the gap to the level of competing methods (the RMG and L2M of GapDIM remain slightly higher than those of Embedding Shift, ReAlign and I0T), we achieve superior retrieval performance, as shown in Tab. 1, which is based on high-importance dimensions. Furthermore, as the

Table 1: Modality-gap, retrieval, and clustering results across backbones and comparison methods.

Model	Method	Flickr30k					MSCOCO					LLaVA-CC3M				
		RMG	L2M	R@1	R@10	V-Meas	RMG	L2M	R@1	R@10	V-Meas	RMG	L2M	R@1	R@10	V-Meas
ViT-B/32	Pretrained	0.80	0.69	56.86	86.56	33.11	0.80	0.74	36.45	**73.32**	61.96	0.80	0.71	**69.74**	93.85	31.18
	Embedding Shift	0.72	0.02	49.13	80.81	**64.35**	0.71	0.01	30.63	65.10	89.19	0.73	0.06	63.59	92.05	**56.46**
	ReAlign	0.72	0.01	44.65	75.45	64.02	0.71	0.02	27.91	61.07	87.88	0.73	0.05	58.46	89.49	54.76
	I0T	0.72	0.02	53.38	83.98	57.68	0.71	0.03	34.64	69.98	**89.23**	0.73	0.06	68.46	93.33	51.02
	GapDIM (ours)	0.78	0.58	**57.11**	**86.61**	36.01	0.77	0.53	**36.67**	73.14	71.07	0.76	0.42	69.72	**93.86**	39.81
	Improvement			*+0.25*	*+0.05*	*+2.90*			*+0.22*	*-0.18*	*+9.12*			*-0.02*	*+0.01*	*+8.63*
ViT-B/16	Pretrained	0.80	0.73	61.26	88.45	37.25	0.81	0.78	39.92	75.34	63.49	0.81	0.76	66.35	**93.08**	29.92
	Embedding Shift	0.71	0.02	53.29	83.50	59.10	0.71	0.00	32.36	67.78	84.54	0.73	0.05	60.86	88.78	37.83
	ReAlign	0.71	0.01	47.70	78.66	56.40	0.71	0.02	29.34	64.20	85.63	0.73	0.04	57.04	86.63	**48.86**
	I0T	0.71	0.02	57.31	86.61	**62.15**	0.71	0.03	36.75	72.38	**90.05**	0.73	0.06	64.92	93.08	42.48
	GapDIM (ours)	0.78	0.58	**61.73**	**88.83**	41.61	0.79	0.67	**40.06**	**75.24**	65.75	0.76	0.43	**66.63**	93.08	35.09
	Improvement			*+0.46*	*+0.38*	*+4.36*			*+0.14*	*-0.10*	*+1.76*			*+0.28*	*-0.08*	*+5.18*
ViT-bigG/14	Pretrained	0.77	0.58	74.27	94.34	38.56	0.77	0.60	46.99	80.80	60.65	0.78	0.58	**74.22**	**94.51**	30.10
	Embedding Shift	0.70	0.03	65.48	90.54	59.40	0.70	0.02	40.12	73.48	86.44	0.72	0.05	69.69	92.84	**58.94**
	ReAlign	0.70	0.01	60.48	86.90	63.49	0.70	0.02	36.23	68.89	86.07	0.72	0.04	64.92	88.07	50.46
	I0T	0.70	0.02	69.21	92.28	**64.85**	0.70	0.02	44.11	77.61	89.91	0.72	0.05	73.75	**94.51**	53.08
	GapDIM (ours)	0.76	0.55	**74.71**	**94.39**	42.82	0.75	0.51	**47.55**	**80.90**	68.56	0.75	0.43	73.75	**94.51**	25.83
	Improvement			*+0.43*	*+0.05*	*+4.26*			*+0.56*	*+0.10*	*+7.91*			*-0.53*	*+0.00*	*-4.27*
ViT-B/16-Sig	Pretrained	0.86	1.00	73.20	94.59	27.88	0.87	1.04	**51.06**	**83.42**	57.98	0.88	1.07	75.26	93.88	29.72
	Embedding Shift	0.71	0.08	64.63	89.84	48.32	0.71	0.03	41.25	74.91	81.02	0.72	0.05	61.48	84.95	**53.57**
	ReAlign	0.71	0.01	56.23	84.25	**53.67**	0.71	0.01	36.81	70.14	87.06	0.72	0.04	58.16	81.38	45.86
	I0T	0.71	0.02	68.91	92.78	53.44	0.71	0.04	46.87	80.11	81.14	0.72	0.08	73.98	94.39	46.73
	GapDIM (ours)	0.77	0.49	**73.59**	**94.94**	33.77	0.85	0.88	50.65	83.09	62.49	0.75	0.33	**75.51**	**94.39**	31.59
	Improvement			*+0.40*	*+0.35*	*+5.88*			*-0.40*	*-0.32*	*+4.51*			*+0.26*	*+0.51*	*+1.87*
ViT-L/16-Sig	Pretrained	0.85	0.97	**77.33**	95.45	37.88	0.86	0.99	51.02	82.13	51.85	0.87	1.03	**72.79**	94.75	24.46
	Embedding Shift	0.71	0.09	69.03	91.95	57.09	0.72	0.06	42.29	74.47	77.88	0.72	0.06	61.58	86.16	**45.69**
	ReAlign	0.71	0.01	59.82	85.88	**63.84**	0.72	0.01	36.27	68.57	**80.17**	0.72	0.04	53.70	81.15	42.19
	I0T	0.71	0.02	73.32	94.42	54.01	0.72	0.03	48.10	80.15	76.10	0.72	0.07	70.88	94.03	44.03
	GapDIM (ours)	0.77	0.53	77.31	**95.57**	38.31	0.77	0.49	**51.22**	**82.57**	60.61	0.76	0.39	**72.79**	**95.23**	28.47
	Improvement			*-0.02*	*+0.12*	*+0.43*			*+0.20*	*+0.44*	*+8.76*			*+0.00*	*+0.48*	*+4.01*

presence of the gap is known to reduce clustering performance Grassucci et al. (2026), our results confirm this trend: removing the noisy gap-driving dimensions usually improves the V-measure. Notably, this improvement occurs even though our method retains a residual gap, suggesting that in high-dimensional spaces semantic groups can organize coherently once the coordinates that scatter embeddings across the space have been removed.

Visualizing pruned coordinates. We inspect the qualitative impact of GapDIM pruned coordinates via the biplot visualization Jolliffe & Cadima (2016); Gabriel (1971). In the biplot, a vector represents a coordinate seen through the first two dimensions of a PCA and, the higher the magnitude of the vector, the more it contributes to those principal components. The red vectors in Fig. 5 correspond to the coordinates pruned by GapDIM, while the green one the retained ones. Interestingly, in the two-dimensional PCA plane, the absolute cosine similarity between the measurable pruned arrows and the projected modality gap direction is much higher than for the retained arrows, around 0.95. This proves that our pruned coordinates are strongly aligned with the projected gap. By removing such coordinates, the two distributions get closer and the retained coordinates span a higher percentage of the space. We therefore show qualitatively that, by pruning these selected coordinates, we improve the multimodal representation space by filtering out gap-carrying coordinates in favor of variance-preserving ones, also observable by the reduction of the explained variance given by the first two PCs. We perform additional and extensive assessments to diagnose the effect of GapDIM on the space in Appendix E.

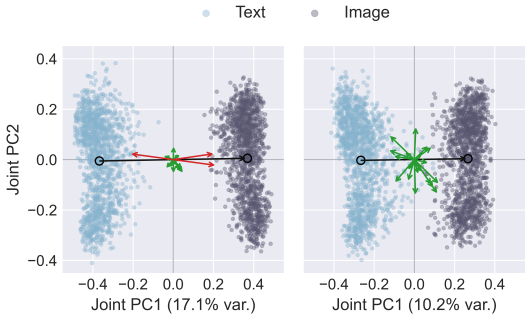

Figure 5: Left: Biplot with coordinates to be pruned by GapDIM (red) that have high cosine similarity with the gap vector. Right: Biplot after coordinate pruning.

Reducing the gap while preserving alignment. Table 1 reports the quantitative results. GapDIM generally matches or slightly surpasses the retrieval performance of the pretrained baseline while improving clustering in most model-dataset configurations. Compared with the evaluated post-hoc methods, it avoids the systematic losses of several R@1 and R@10 points observed when the complete embedding distribution is transformed. Removing fewer than 3% of the coordinates usually

preserves retrieval performance and improves clustering, indicating a favorable gap-reduction versus representation-preservation trade-off.

Figure 1 shows the organization of the latent space before (left, pretrained) and after applying GapDIM (right) on the CIFAR-10 dataset. In the left part of the figure, the modality gap is clearly visible. When applying the proposed GapDIM, the gap is reduced, cross-modal alignment is preserved, and semantic groups become visibly more compact (accuracy also increases from 72.12 to 100.0), a direct consequence of removing the noisy, gap-driving coordinates.

Cross-modal generation. We evaluate image-free zero-shot image captioning to test whether the representation learned from one modality can be used by a decoder operating on the other one. Experimental details in Appendix I. By applying GapDIM on the pretrained captioning model, we improve the performance in all the considered metrics. This provides evidence that pruning the selected coordinates positively affects the results in the captioning task too.

Table 2: Zero-shot captioning on MSCOCO with CLIP ViT-B/32.

Method	BLEU-1	BLEU-4	CIDEr	SPICE
Pretrained	53.73	12.54	32.54	10.28
GapDIM	**55.13**	**13.31**	**40.37**	**11.77**

Ablation studies. We evaluate single-threshold variants on MSCOCO to disentangle the effects of the two pruning criteria. As shown in Table 3 (top), using both thresholds allows GapDIM to match the retrieval performance of the pretrained model while substantially improving clustering. Applying only τ_{gap} yields a larger clustering improvement but degrades retrieval, whereas applying only τ_{imp} better preserves retrieval but provides a smaller clustering improvement.

We further ablate the number of samples used to estimate the SVD subspaces. As shown in Table 3, GapDIM is robust to the size of the support set used for subspace estimation and subsequent coordinate pruning. Using as few as 1k samples recovers the retrieval and clustering performance obtained with larger support sets, including the complete dataset. This indicates that the pruning criterion does not require a large calibration set.

Table 3: Ablation studies on the single-threshold scenario versus the pretrained model and the complete GapDIM criterion (top), and effect of the number of samples used to estimate D_{sub} (bottom).

Method	RMG	L2M	R@1	V-Meas
Pretrained	0.80	0.74	<u>36.45</u>	61.96
τ_{gap} only	0.73	0.34	33.26	**85.92**
τ_{imp} only	0.76	0.51	35.84	64.47
GapDIM (ours)	0.77	0.53	**36.67**	<u>71.07</u>
N_{fit}				
Pretrained	0.77	0.60	46.98	60.65
100	0.76	0.40	26.15	50.47
1k	0.73	0.50	47.55	68.56
10k	0.75	0.50	47.55	68.56
100k	0.75	0.50	47.55	68.56
All	0.75	0.50	47.55	68.56

We report additional experiments on a synthetic dataset and the evaluation on the preservation of unimodal distributions in Appendices G and I, respectively. Furthermore, we report additional ablations in Appendix F, which include cross-dataset and cross-model threshold transfer to account for zero-shot distribution shift, among others.

5 Conclusion

In this paper, we investigated the relationship between the modality gap in contrastive VLMs and the semantic structure of the shared embedding space. We showed that the dimensions responsible for the gap are consistently those with the lowest importance for the shared principal subspace, and that no dimension simultaneously carries high gap and high semantic relevance. Building on this, we proposed GapDIM, a post-hoc, training-free method that identifies and prunes this small subset of noisy, gap-driving dimensions. This targeted removal does not require any update to the encoder parameters and can be applied directly to any pretrained contrastive model, including CLIP and SigLIP variants of different sizes. Our empirical validation demonstrated that GapDIM matches or surpasses the retrieval performance of the pretrained baseline, while consistently improving clustering across all tested configurations, opening new directions for the deployment of multimodal representations in tasks that jointly require pairwise alignment and global semantic coherence.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance, i.e., polishing and proofreading the text (grammar and formatting) and for discussing the research ideas we have previously developed. Retrieval of related work and literature review were conducted manually by the authors. All technical claims, experimental results, and artifacts were produced by the authors. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

We describe GapDIM in full algorithmic detail in Sec. 3.2, including every step from the raw paired embeddings to the final retained coordinate set; the method requires no encoder training and involves only centroid computation, SVD, and threshold-based coordinate selection. Section D.1 details the exact threshold-search procedure. All evaluation metrics are formally defined in Appendix H, and all datasets used (CIFAR-10, Flickr30k, MSCOCO, and LLaVA-CC3M) are public benchmarks evaluated on their official or standard splits. All experiments, including every ablation and evaluation reported in this appendix, were run on a single NVIDIA A6000 GPU with 48GB of VRAM, and results reported with three random seeds. Anonymized code implementing GapDIM and all experiments in this paper is available at the link given in the abstract.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Neural Information Processing Systems (NeurIPS)*, 2022.

Na Min An, Eunki Kim, James Thorne, and Hyunjung Shim. I0T: Embedding standardization method towards zero modality gap. In *ACL Long Papers*, pp. 27182–27199, 2025.

Rhea Chowers, Oshri Naparstek, Udi Barzelay, and Yair Weiss. Is the modality gap a bug or a feature? a robustness perspective. *ArXiv preprint: arXiv:2603.29080*, 2026.

Antonio D'Orazio, Maria Rosaria Briglia, Donato Crisostomi, Dario Loi, Emanuele Rodolà, and Iacopo Masi. Implicit inversion turns CLIP into a decoder. In *International Conference on Learning Representations (ICLR)*, 2026.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *ArXiv preprint: arXiv:1109.3843*, 2012.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in clip. *International Conference on Learning Representations (ICLR)*, 2025.

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

K. R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971. doi: 10.1093/biomet/58.3.453.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. CyCLIP: Cyclic contrastive language-image pretraining. In *Neural Information Processing Systems (NeurIPS)*, 2022.

Eleonora Grassucci, Giordano Cicchetti, and Danilo Comminiello. Closing the modality gap enables novel multimodal learning applications. In *Workshop on Representational Alignment at ICLR 2025*, 2025.

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

Jinjin Guo, Yexin Li, Zhichao Huang, Jun Fang, Zhiyuan Liu, Chao Liu, Pengzhang Liu, and Qixia Jiang. Spectral disentanglement and enhancement: A dual-domain contrastive framework for representation learning. In *ACM Web Conference 2026*, pp. 4127–4136, 2026.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011.

Eric Hartford, Lucas Atkins, Fernando Fernandes Neto, , and David Golchinfar. Spectrum: Targeted training on signal to noise ratio. *ArXiv preprint: arXiv:2406.06623*, 2024.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Deciphering cross-modal alignment in large vision-language models with modality integration rate. In *International Conference on Computer Vision (ICCV)*, 2025.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 2009.

Jeong Ryong Lee, Yejee Shin, Geonhui Son, and Dosik Hwang. Diffusion bridge: Leveraging diffusion model to reduce the modality gap between text and vision for zero-shot image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Binxu Li, Yuhui Zhang, Xiaohan Wang, Weixin Liang, Ludwig Schmidt, and Serena Yeung-Levy. Closing the modality gap for mixed modality search. *ArXiv preprint: arXiv:2507.19054*, 2025.

Jiaoyang Li, Jun Fang, Tianhao Gao, Xiaohui Zhang, Zhiyuan Liu, Chao Liu, Pengzhang Liu, and Qixia Jiang. FANoise: Singular value-adaptive noise modulation for robust multimodal representation learning. *AAAI Conference on Artificial Intelligence*, 40(8):6199–6207, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Neural Information Processing Systems (NeurIPS)*, 2022.

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

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

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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 (CVPR)*, 2026.

Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: Clip prefix for image captioning. *ArXiv preprint: arXiv:2111.09734*, 2021.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019. URL `https://cdn.openai.com/better-language-models/language-models.pdf`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27253–27262, 2024.

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

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

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

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv preprint: arXiv:1807.03748*, 2018.

Yimu Wang, Evelien Riddell, Adrian Chow, Sean Sedwards, and Krzysztof Czarnecki. Mitigating the modality gap: Few-shot out-of-distribution detection with multi-modal prototypes and image bias estimation. In *Winter Conference on Applications of Computer Vision (WACV)*, 2026.

Zhongxing Xu, Feilong Tang, Zhe Chen, Yingxue Su, Zhiyi Zhao, Ge Zhang, Jionglong Su, and Zongyuan Ge. Toward modality gap: Vision prototype learning for weakly-supervised semantic segmentation with CLIP. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 9023–9031, 2025.

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

Xiaomin Yu, Yi Xin, Wenjie Zhang, Chonghan Liu, Hanzhen Zhao, Xiaoxing Hu, Xinlei Yu, Ziyue Qiao, Hao Tang, Xue Yang, Xiaobin Hu, Chengwei Qin, Hui Xiong, Yu Qiao, and Shuicheng Yan. Modality gap-driven subspace alignment training paradigm for multimodal large language models. *ArXiv preprint: arXiv:2602.07026*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952, October 2023.

Yuhui Zhang, Jeff Z. HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. In *International Conference on Learning Representations (ICLR)*, 2023.

Yuhui Zhang, Elaine Sui, and Serena Yeung-Levy. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. In *International Conference on Learning Representations (ICLR)*, 2024.

A Complete Proofs

This appendix provides the complete derivations of the theoretical results stated in Section 3.2. We use the modality centroids $\boldsymbol{\mu}_X, \boldsymbol{\mu}_Y$ defined in equation 3, the dominant centered-subspace bases $\mathbf{W}_M$ defined in equation 7, and the canonical basis vectors $\mathbf{e}_j$ introduced in equation 8.

To avoid ambiguity between the raw centroid displacement used in the proofs and the normalized gap score used by the pruning rule, we introduce the following proof-specific notation. Let

$$\boldsymbol{\delta} = \boldsymbol{\mu}_X - \boldsymbol{\mu}_Y \tag{15}$$

denote the signed centroid-gap vector, and let

$$d_j = |\delta_j| = |\mu_{X,j} - \mu_{Y,j}| \tag{16}$$

denote its raw magnitude along coordinate j.

The operational pruning score $g_j \in [0, 1]$ is obtained by min–max normalizing the raw magnitudes d_j; this operation preserve the ranking of dimensions with respect to the gap value each coordinate carries.

We assume throughout that $\boldsymbol{\delta} \neq \mathbf{0}$ and define

$$a_j = \frac{d_j^2}{\|\boldsymbol{\delta}\|_2^2} = \frac{\delta_j^2}{\|\boldsymbol{\delta}\|_2^2}, \qquad \sum_{j=1}^{D} a_j = 1, \tag{17}$$

together with the normalized gap direction

$$\mathbf{u} = \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2}. \tag{18}$$

For each modality $M \in \{X, Y\}$, we define the orthogonal projector

$$\mathbf{P}_M = \mathbf{W}_M \mathbf{W}_M^\top \tag{19}$$

and use

$$\ell_{M,j} = \mathbf{e}_j^\top \mathbf{P}_M \mathbf{e}_j, \qquad \rho_M = \|\mathbf{P}_M \mathbf{u}\|_2^2. \tag{20}$$

A.1 Preliminary Lemma: Properties of the Subspace Projector

Lemma 1. For each modality $M \in \{X, Y\}$, the matrix $\mathbf{P}_M = \mathbf{W}_M \mathbf{W}_M^\top$ is an orthogonal projector: it is symmetric and idempotent,

$$\mathbf{P}_M^\top = \mathbf{P}_M, \qquad \mathbf{P}_M^2 = \mathbf{P}_M. \tag{21}$$

Moreover, for every unit vector $\mathbf{v} \in \mathbb{R}^D$,

$$0 \leq \mathbf{v}^\top \mathbf{P}_M \mathbf{v} \leq 1. \tag{22}$$

In particular, $\ell_{M,j} \in [0, 1]$ for every coordinate j, and $\rho_M \in [0, 1]$.

Proof. Since the columns of $\mathbf{W}_M$ are orthonormal,

$$\mathbf{W}_M^\top \mathbf{W}_M = \mathbf{I}_{D_{\text{sub}}}. \tag{23}$$

Symmetry follows directly from

$$\mathbf{P}_M^\top = (\mathbf{W}_M \mathbf{W}_M^\top)^\top = \mathbf{W}_M \mathbf{W}_M^\top = \mathbf{P}_M. \tag{24}$$

Similarly,

$$\mathbf{P}_M^2 = \mathbf{W}_M (\mathbf{W}_M^\top \mathbf{W}_M) \mathbf{W}_M^\top = \mathbf{W}_M \mathbf{I}_{D_{\text{sub}}} \mathbf{W}_M^\top = \mathbf{P}_M. \tag{25}$$

For any $\mathbf{v} \in \mathbb{R}^D$, symmetry and idempotence give

$$\mathbf{v}^\top \mathbf{P}_M \mathbf{v} = \mathbf{v}^\top \mathbf{P}_M^\top \mathbf{P}_M \mathbf{v} = \|\mathbf{P}_M \mathbf{v}\|_2^2 \geq 0. \tag{26}$$

The complementary matrix $\mathbf{I}_D - \mathbf{P}_M$ is also symmetric. Moreover,

$$(\mathbf{I}_D - \mathbf{P}_M)^2 = \mathbf{I}_D - 2\mathbf{P}_M + \mathbf{P}_M^2 = \mathbf{I}_D - \mathbf{P}_M, \tag{27}$$

where the last equality follows from equation 25. Therefore, $\mathbf{I}_D - \mathbf{P}_M$ is also an orthogonal projector, and

$$\|\mathbf{v}\|_2^2 - \mathbf{v}^\top \mathbf{P}_M \mathbf{v} = \mathbf{v}^\top (\mathbf{I}_D - \mathbf{P}_M)\mathbf{v} = \|(\mathbf{I}_D - \mathbf{P}_M)\mathbf{v}\|_2^2 \geq 0. \tag{28}$$

Consequently,

$$\mathbf{v}^\top \mathbf{P}_M \mathbf{v} \leq \|\mathbf{v}\|_2^2. \tag{29}$$

For $\|\mathbf{v}\|_2 = 1$, this proves equation 22.

Finally, since $\|\mathbf{e}_j\|_2 = 1$, taking $\mathbf{v} = \mathbf{e}_j$ gives

$$\ell_{M,j} = \mathbf{e}_j^\top \mathbf{P}_M \mathbf{e}_j \in [0, 1]. \tag{30}$$

Likewise, $\|\mathbf{u}\|_2 = 1$ by construction, and therefore

$$\rho_M = \|\mathbf{P}_M \mathbf{u}\|_2^2 = \mathbf{u}^\top \mathbf{P}_M \mathbf{u} \in [0, 1]. \tag{31}$$

A.2 Proof of Proposition 2 (Gap-Leverage Bound)

Proposition 2. For every modality $M \in \{X, Y\}$ and coordinate $j \in \{1, \ldots, D\}$,

$$\sqrt{a_j} \;\leq\; \sqrt{\ell_{M,j}\,\rho_M} + \sqrt{(1 - \ell_{M,j})(1 - \rho_M)}. \tag{32}$$

Proof. **Step 1 (reduction to an inner product).** Since $\mathbf{u} = \mathbf{g}/\|\mathbf{g}\|_2$,

$$\mathbf{e}_j^\top \mathbf{u} = \frac{\mathbf{e}_j^\top \mathbf{g}}{\|\mathbf{g}\|_2} = \frac{g_j}{\|\mathbf{g}\|_2}, \qquad \text{so} \qquad (\mathbf{e}_j^\top \mathbf{u})^2 = \frac{g_j^2}{\|\mathbf{g}\|_2^2} = a_j. \tag{33}$$

Hence $\sqrt{a_j} = |\mathbf{e}_j^\top \mathbf{u}|$, and it suffices to bound $|\mathbf{e}_j^\top \mathbf{u}|$.

Step 2 (orthogonal decomposition). Because $\mathbf{P}_M$ is idempotent and symmetric (Lemma 1), $\mathbf{I} - \mathbf{P}_M$ is the complementary orthogonal projector, and every vector decomposes uniquely as

$$\mathbf{e}_j = \underbrace{\mathbf{P}_M \mathbf{e}_j}_{\mathbf{e}_j^{\parallel}} + \underbrace{(\mathbf{I} - \mathbf{P}_M)\mathbf{e}_j}_{\mathbf{e}_j^{\perp}}, \qquad \mathbf{u} = \underbrace{\mathbf{P}_M \mathbf{u}}_{\mathbf{u}^{\parallel}} + \underbrace{(\mathbf{I} - \mathbf{P}_M)\mathbf{u}}_{\mathbf{u}^{\perp}}. \tag{34}$$

Step 3 (the cross terms vanish). Expanding the inner product,

$$\mathbf{e}_j^\top \mathbf{u} = (\mathbf{e}_j^{\parallel})^\top \mathbf{u}^{\parallel} + (\mathbf{e}_j^{\parallel})^\top \mathbf{u}^{\perp} + (\mathbf{e}_j^{\perp})^\top \mathbf{u}^{\parallel} + (\mathbf{e}_j^{\perp})^\top \mathbf{u}^{\perp}. \tag{35}$$

For the second term, using $\mathbf{P}_M(\mathbf{I} - \mathbf{P}_M) = \mathbf{P}_M - \mathbf{P}_M^2 = \mathbf{0}$ (equation 25) and the symmetry of $\mathbf{P}_M$,

$$(\mathbf{e}_j^{\parallel})^\top \mathbf{u}^{\perp} = \mathbf{e}_j^\top \mathbf{P}_M^\top (\mathbf{I} - \mathbf{P}_M)\mathbf{u} = \mathbf{e}_j^\top \mathbf{P}_M (\mathbf{I} - \mathbf{P}_M)\mathbf{u} = 0, \tag{36}$$

and, symmetrically, $(\mathbf{e}_j^{\perp})^\top \mathbf{u}^{\parallel} = \mathbf{e}_j^\top (\mathbf{I} - \mathbf{P}_M)\mathbf{P}_M \mathbf{u} = \mathbf{e}_j^\top (\mathbf{P}_M - \mathbf{P}_M^2)\mathbf{u} = 0$. Substituting into equation 35,

$$\mathbf{e}_j^\top \mathbf{u} = (\mathbf{e}_j^{\parallel})^\top \mathbf{u}^{\parallel} + (\mathbf{e}_j^{\perp})^\top \mathbf{u}^{\perp}. \tag{37}$$

Step 4 (triangle inequality and Cauchy–Schwarz). From equation 37, we have

$$\mathbf{e}_j^\top \mathbf{u} = (\mathbf{e}_j^{\parallel})^\top \mathbf{u}^{\parallel} + (\mathbf{e}_j^{\perp})^\top \mathbf{u}^{\perp}. \tag{38}$$

Taking the absolute value of both sides gives

$$|\mathbf{e}_j^\top \mathbf{u}| = \left| (\mathbf{e}_j^{\parallel})^\top \mathbf{u}^{\parallel} + (\mathbf{e}_j^{\perp})^\top \mathbf{u}^{\perp} \right|. \tag{39}$$

Applying the triangle inequality,

$$|\mathbf{e}_j^\top \mathbf{u}| \leq \left| (\mathbf{e}_j^{\parallel})^\top \mathbf{u}^{\parallel} \right| + \left| (\mathbf{e}_j^{\perp})^\top \mathbf{u}^{\perp} \right|. \tag{40}$$

For any two vectors $\mathbf{x}$ and $\mathbf{y}$, the Cauchy–Schwarz inequality gives

$$|\mathbf{x}^\top \mathbf{y}| \le \|\mathbf{x}\|_2 \|\mathbf{y}\|_2. \tag{41}$$

Applying it separately to the parallel components,

$$\left|(\mathbf{e}_j^{\|})^\top \mathbf{u}^{\|}\right| \le \|\mathbf{e}_j^{\|}\|_2 \|\mathbf{u}^{\|}\|_2, \tag{42}$$

and to the orthogonal components,

$$\left|(\mathbf{e}_j^{\perp})^\top \mathbf{u}^{\perp}\right| \le \|\mathbf{e}_j^{\perp}\|_2 \|\mathbf{u}^{\perp}\|_2. \tag{43}$$

Combining these two bounds yields

$$|\mathbf{e}_j^\top \mathbf{u}| \le \|\mathbf{e}_j^{\|}\|_2 \|\mathbf{u}^{\|}\|_2 + \|\mathbf{e}_j^{\perp}\|_2 \|\mathbf{u}^{\perp}\|_2. \tag{44}$$

Step 5 (identifying the norms). By the same argument as in equation 26,

$$\|\mathbf{e}_j^{\|}\|_2^2 = \mathbf{e}_j^\top \mathbf{P}_M \mathbf{e}_j = \ell_{M,j}, \qquad \|\mathbf{u}^{\|}\|_2^2 = \mathbf{u}^\top \mathbf{P}_M \mathbf{u} = \rho_M, \tag{45}$$

and, since $\|\mathbf{e}_j\|_2 = \|\mathbf{u}\|_2 = 1$, equation 28 gives

$$\|\mathbf{e}_j^{\perp}\|_2^2 = 1 - \ell_{M,j}, \qquad \|\mathbf{u}^{\perp}\|_2^2 = 1 - \rho_M. \tag{46}$$

Substituting equation 45-equation 46 into equation 44 and using $\sqrt{a_j} = |\mathbf{e}_j^\top \mathbf{u}|$ from Step 1 yields exactly equation 32, which is precisely equation 14.

Remark (equality conditions). Equality requires the two projected vector pairs to be linearly dependent, namely $\mathbf{u}^{\|} = c_{\|}\mathbf{e}_j^{\|}$ and $\mathbf{u}^{\perp} = c_{\perp}\mathbf{e}_j^{\perp}$, together with $c_{\|}c_{\perp} \ge 0$, so that the two inner products have the same sign. These conditions are required simultaneously for equality in both applications of Cauchy–Schwarz and in the triangle inequality.

A.3 Proof of Corollary 1 (Perfect-Orthogonality Regime)

The following corollary formalizes the perfect-orthogonality case discussed after equation 14 in the main text.

Corollary 1. If $\mathbf{P}_M \boldsymbol{\delta} = \mathbf{0}$, then $a_j \le 1 - \ell_{M,j}$ for every coordinate j.

Proof. Since $\mathbf{u}$ is a scalar multiple of $\boldsymbol{\delta}$ by equation 18, $\mathbf{P}_M \boldsymbol{\delta} = \mathbf{0}$ implies $\mathbf{P}_M \mathbf{u} = \mathbf{0}$, so $\rho_M = \|\mathbf{P}_M \mathbf{u}\|_2^2 = 0$. Substituting $\rho_M = 0$ into equation 32,

$$\sqrt{a_j} \;\le\; \sqrt{\ell_{M,j} \cdot 0} + \sqrt{(1 - \ell_{M,j}) \cdot 1} = \sqrt{1 - \ell_{M,j}}. \tag{47}$$

Squaring both sides, which are non-negative by Lemma 1, gives $a_j \le 1 - \ell_{M,j}$. This is precisely the perfect-orthogonality specialization discussed after equation 14 in the main text.

B Mechanistic Origin of the Gap-Leverage Pattern

Proposition 1 (Section 3.3) shows that a modality-wide translation $\mathbf{c}_M$ changes the gap vector by $\mathbf{c}_X - \mathbf{c}_Y$ while leaving the centered subspaces and the leverage scores $\ell_{M,j}$ exactly unchanged. This establishes that the two quantities *can* decouple, but it does not explain why trained contrastive encoders would produce embeddings in which such a component is actually present, this is done in Liang et al. (2022), Schrodi et al. (2025). This appendix summarizes the training-dynamics mechanism that motivates this regime and states its precise quantitative consequence through Corollary 1.

Why a modality-constant offset can survive contrastive training. Zhang et al. (2024) model the difference between paired representations as $\mathbf{x}_i - \mathbf{y}_i = \mathbf{c}_\perp + \boldsymbol{\epsilon}_i$ (equation 3.3), where $\mathbf{c}_\perp$ is an approximately constant modality-dependent component and $\boldsymbol{\epsilon}_i$ is sample-dependent alignment noise. Their analysis of the symmetric contrastive gradient (their Lemma 1) shows that, for the

InfoNCE-style objective in equation 1, every gradient term is built from *differences* between embeddings belonging to the same modality (an anchor against its in-batch negatives). A component that is added identically to every embedding of a modality, and that is orthogonal to the span of the centered embeddings of that modality, cancels from all such within-modality differences and therefore receives zero gradient. Because contrastive pretraining typically starts from two independently initialized encoders, a modality-specific constant component present at initialization is not guaranteed to be corrected during training if it satisfies this orthogonality condition; it can instead persist throughout optimization. Related gradient-flow analyses of contrastive multimodal training reach a consistent conclusion: a nonzero modality gap can be created and perpetuated purely by the optimization dynamics of the contrastive objective, without requiring any explicit gap-inducing term in the loss (Yaras et al., 2025). A closely related empirical account of directions that are approximately constant within a modality but differ across modalities is given by Zhang et al. (2023).

Connecting the mechanism to the gap-leverage bound. The mechanism above corresponds precisely to the orthogonality model $\mathbf{P}_M \mathbf{g} = \mathbf{0}$ used in Corollary 1 (Appendix A.3): $\mathbf{c}_\perp$ being orthogonal to the centered span of modality M means that the gap direction it induces has zero projection onto $\mathbf{P}_M$. Under this model, Corollary 1 gives

$$a_j \;\le\; 1 - \ell_{M,j} \qquad \text{for every coordinate } j, \tag{48}$$

i.e., a coordinate that carries a large share of the modality-gap energy *must* have low leverage in the dominant centered subspace of M. This is the quantitative prediction that GapDIM exploits: coordinates dominated by a persistent modality-constant offset are, by this mechanism, exactly the coordinates with low subspace participation, which is the high-gap/low-leverage pattern reported empirically throughout Section 4 and Fig. 4.

Scope of the mechanism. This training-dynamics explains why a high-gap, low-leverage coordinate pattern is a plausible outcome of contrastive pretraining; it does not by itself guarantee that $\rho_M \approx 0$ for any specific trained model, since the orthogonality condition is only approximate in practice. Whether real pretrained VLMs actually operate in a regime where the general bound of Proposition 2 (Section 3.3) is informative is an empirical question, which we address quantitatively in Appendix C.

C Empirical Validation of the Gap-Leverage Bound

Proposition 2 (Section 3.3) holds for any value of the subspace overlap $\rho_M \in [0, 1]$, but its usefulness as an explanation depends on how close real pretrained VLMs are to the orthogonal regime $\rho_M \approx 0$, and on how tight the bound is in practice. We quantify both on MSCOCO with CLIP ViT-B/32 ($D = 512$, $D_{\text{sub}} = 256$), fitting $\mathbf{W}_X, \mathbf{W}_Y$ on the training split and evaluating all quantities below on the independent validation split.

C.1 Subspace Overlap

We estimate the subspace overlap directly from equation 13, obtaining

$$\rho_X = 0.635, \qquad \rho_Y = 0.575, \tag{49}$$

for the text and image dominant centered subspaces, respectively. We report 95% confidence intervals obtained by bootstrap: we resample the validation pairs with replacement 1,000 times, recompute $\mathbf{g}$ and ρ_M on each resample, and take the 2.5th and 97.5th percentiles of the resulting distribution. This gives $\rho_X \in [0.63, 0.64]$ and $\rho_Y \in [0.57, 0.58]$. Both intervals are far from zero, indicating that exact orthogonality between the gap direction and the dominant centered subspaces is not observed in this configuration; the general bound of Proposition 2, rather than the $\rho_M = 0$ special case of Corollary 1, is the operative regime for real pretrained embeddings.

C.2 Tightness of the Bound

For every modality M and coordinate j, we compute the tightness ratio

$$\tau_{M,j} = \frac{a_j}{\left[\sqrt{\ell_{M,j}\rho_M} + \sqrt{(1-\ell_{M,j})(1-\rho_M)}\right]^2} \;\in [0, 1], \tag{50}$$

which measures how close the observed normalized gap energy a_j comes to saturating the right-hand side of equation 32. The maximum value attained across both modalities and all $D = 512$ coordinates is

$$\max_{M \in \{X,Y\},\, j \in \{1,\dots,D\}} \tau_{M,j} = 0.194. \tag{51}$$

Even the single coordinate closest to the theoretical envelope reaches only 19.4% of its allowed upper bound. This is consistent with the equality-condition remark in Appendix A.2: exact equality requires a non-generic colinearity between $\mathbf{e}_j$ and $\mathbf{u}$ inside and outside the subspace simultaneously, which is not expected to occur in a 512-dimensional embedding space. We therefore present Proposition 2 as a valid worst-case geometric guarantee rather than as a quantity that real coordinate-level gap allocations approach.

C.3 Benefit-Cost Concentration of the GapDIM Removal Set

The bound bounds individual coordinates; we now quantify the aggregate benefit-cost trade-off for the coordinate set that GapDIM actually removes. Let $\mathcal{R} = \{70, 152, 253, 285, 307\}$ be the five coordinates removed by GapDIM in this configuration (Table 1), and let $\mathbf{D}_{\mathcal{R}} \in \{0,1\}^{D \times D}$ be the diagonal matrix with $(\mathbf{D}_{\mathcal{R}})_{jj} = 1$ if $j \in \mathcal{R}$ and 0 otherwise. Since $\mathbf{D}_{\mathcal{R}}$ is a diagonal $0/1$ projector onto the coordinates in $\mathcal{R}$,

$$\|\mathbf{g}\|_2^2 - \|(\mathbf{I} - \mathbf{D}_{\mathcal{R}})\mathbf{g}\|_2^2 = \sum_{j \in \mathcal{R}} g_j^2, \qquad \|\mathbf{D}_{\mathcal{R}}\mathbf{W}_M\|_F^2 = \sum_{j \in \mathcal{R}} \ell_{M,j}, \tag{52}$$

where the first identity follows because $(\mathbf{I} - \mathbf{D}_{\mathcal{R}})\mathbf{g}$ retains exactly the coordinates of $\mathbf{g}$ outside $\mathcal{R}$, and the second because $\mathbf{D}_{\mathcal{R}}\mathbf{W}_M$ retains exactly the rows of $\mathbf{W}_M$ indexed by $\mathcal{R}$, whose squared Frobenius norm is $\sum_{j \in \mathcal{R}} \sum_{k=1}^{D_{\text{sub}}} (\mathbf{W}_M)_{j,k}^2 = \sum_{j \in \mathcal{R}} \ell_{M,j}$ by the definition of $\ell_{M,j}$ in equation 8. The removed gap energy is therefore the left-hand side of equation 52 normalized by $\|\mathbf{g}\|_2^2$, and the removed leverage mass is $\sum_{j=1}^{D} \ell_{M,j} = \operatorname{tr}(\mathbf{P}_M) = D_{\text{sub}}$, by equation 8 and the orthonormality of $\mathbf{W}_M$. Evaluating these quantities gives

$$\frac{\sum_{j \in \mathcal{R}} g_j^2}{\sum_{j=1}^{D} g_j^2} = 47.70\%, \qquad 100 \cdot \frac{\sum_{j \in \mathcal{R}} \ell_{X,j}}{D_{\text{sub}}} = 0.560\%, \qquad 100 \cdot \frac{\sum_{j \in \mathcal{R}} \ell_{Y,j}}{D_{\text{sub}}} = 0.467\%. \tag{53}$$

In words: the five coordinates GapDIM removes in this configuration (fewer than 1% of $D = 512$) concentrate nearly half of the total modality gap, while touching less than 0.6% of the dominant subspace participation mass in either modality. This is the empirical counterpart of the rationale underlying the pruning rule in Section 3.2: a_j measures the benefit for reducing the gap of removing coordinate j, while $\ell_{M,j}$ its participation cost, in the dominant subspace.

C.4 Summary

Table 4 collects the three measurements. Together they show that (i) the gap is not exactly orthogonal to the dominant centered subspaces, so the general bound of Proposition 2 rather than its orthogonal special case is the relevant regime; (ii) the bound is a loose worst-case envelope rather than a tight approximation, consistent with the high dimensionality of the embedding space; and (iii) despite the bound being loose in the worst case, the coordinates GapDIM actually selects for removal realize a highly favorable empirical benefit-cost trade-off.

D Threshold Selection

Section 3.2 introduces the GapDIM calibration procedure, which selects $(\tau_{\text{imp}}, \tau_{\text{gap}})$ by maximizing calibration-split R@1. We now detail the calibration procedure and propose an unsupervised alternative that can be used as a default setting to avoid calibration.

D.1 TPE Search Procedure

We optimize τ_{imp} and τ_{gap} using Optuna with a `TPESampler`. The Tree-structured Parzen Estimator (TPE) is a sequential model-based optimization strategy that, at each trial, models promising

Table 4: Summary of the empirical validation of the gap-leverage bound on MSCOCO with CLIP ViT-B/32 ($D = 512$, $D_{\text{sub}} = 256$).

Quantity	Value
ρ_X (text), 95% bootstrap CI	$0.635, [0.63, 0.64]$
ρ_Y (image), 95% bootstrap CI	$0.575, [0.57, 0.58]$
Maximum tightness ratio $\max \tau_{M,j}$	0.194
Removed gap energy, $\mathcal{R} = \{70, 152, 253, 285, 307\}$	47.70%
Removed text leverage mass	0.560%
Removed image leverage mass	0.467%

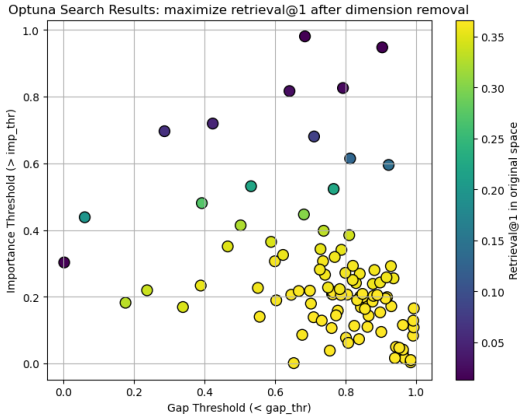

Figure 6: Tree-structured Parzen Estimation over the thresholds τ_{imp} and τ_{gap} naturally converges to cutting dimensions with low subspace participation and high gap: the best trials concentrate in the low-importance, high-gap region. Model is ViT-B/32 (LAION-2B) on MSCOCO.

and non-promising regions of the search space separately and proposes new candidates by favoring values with a high ratio $l(x)/g(x)$, where $l(x)$ is fitted on the best observed trials and $g(x)$ on the remaining ones. We run the search for 100 trials. In each trial, the objective samples a candidate pair $(\tau_{\text{imp}}, \tau_{\text{gap}})$, computes the retained set $\mathcal{I}$, and evaluates R@1_{cal} on the filtered calibration embeddings; the best trial identifies the threshold pair used for the frozen mask. This procedure avoids exhaustive grid search over the two-dimensional threshold space, remains fully training-free since no encoder parameter is updated, and performs hyperparameter selection only over the post-hoc filtering rule applied to the embedding space.

As an example, Figure 6 visualizes the outcome of the search on MSCOCO with ViT-B/32. Trials naturally concentrate in the low-importance, high-gap region of the $(\tau_{\text{imp}}, \tau_{\text{gap}})$ space, confirming that maximizing retrieval consistently rediscovers the high-gap/low-leverage removal criterion motivated theoretically in Section 3. This behavior is consistent across all evaluated model-dataset configurations.

D.2 Unsupervised Distributional Threshold Rule

The TPE search above uses R@1 feedback on a paired calibration set. We also consider a fully unsupervised alternative that requires neither pair identities nor retrieval feedback, using only the geometry of the coordinate statistics (ℓ_j, g_j) defined in equation 9 and equation 4. Figure 4 shows that the high-gap coordinates form a sparse, low-participation region, while the high-gap/high-participation quadrant is empty. Interestingly, such a two-component structure can be recovered by a density model. Therefore, it is possible to fit a two-component Gaussian mixture model (GMM) over the points $\{(\ell_j, g_j)\}_{j=1}^{D}$, and identify the component with the lower mean importance and higher mean gap, and then prune coordinates assigned to that component with posterior probability above the 99-th percentile of the fitted mixture. This replaces the calibration-based threshold search with an unsupervised rule derived entirely from the embedding geometry.

Table 5: Unsupervised distributional threshold rule (two-component GMM on (ℓ_j, g_j), 99th-percentile posterior cutoff) versus the default TPE-tuned thresholds, on MSCOCO.

Model	Method	R@1	R@10	V-measure
ViT-B/32	Distributional fit	34.88	71.25	70.81
	GapDIM (TPE-tuned)	**36.67**	**73.14**	**71.07**
ViT-L/16-SigLIP2-256	Distributional fit	50.04	81.01	54.44
	GapDIM (TPE-tuned)	**51.22**	**82.57**	**60.61**

Table 5 compares this distributional rule with the default TPE-tuned thresholds on MSCOCO, for both a CLIP backbone (ViT-B/32) and a SigLIP2 backbone (ViT-L/16-SigLIP2-256). The distributional rule is consistently comparable to, even though slightly below, the TPE-tuned thresholds on both backbones. We recommend it as a general-purpose, calibration-free fallback when paired retrieval labels are unavailable at deployment time, while the proposed retrieval-tuned procedure remains the best-performing option whenever a small paired calibration set can be obtained.

E Low-Rank and Subspace Assessment

This appendix reports complementary assessment of the low-rank structure underlying GapDIM: how much coordinate variance is retained at the GapDIM stopping point, how retrieval and clustering behave as the subspace rank D_{sub} varies, two multimodal Eckart-Young controls that directly probe whether the modality gap is a low-rank, exposable direction. All evaluations use MSCOCO with ViT-B/32 unless otherwise noted.

E.1 Cumulative Coordinate Variance Diagnostic

We measure how the retained coordinate variance changes as a function of the number of coordinates kept according to the GapDIM ranking (most retained coordinates first). Given the normalized embeddings $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{N\times D}$, let

$$\mathbf{Z} = \begin{bmatrix}\mathbf{X}\\ \mathbf{Y}\end{bmatrix} \in \mathbb{R}^{2N\times D}, \qquad \mu_j = \frac{1}{2N}\sum_{i=1}^{2N} Z_{ij}, \tag{54}$$

be the row-stacked joint matrix and its coordinate-wise mean. The joint coordinate variance of coordinate j is

$$v_j^{\text{joint}} = \frac{1}{2N-1}\sum_{i=1}^{2N}(Z_{ij}-\mu_j)^2. \tag{55}$$

Let π order coordinates by the GapDIM ranking, with retained coordinates first and pruned coordinates last. The cumulative retained variance after keeping the first m coordinates in this order is

$$\text{CEV}_{\text{joint}}(m) = \frac{\sum_{\ell=1}^{m} v_{\pi_\ell}^{\text{joint}}}{\sum_{j=1}^{D} v_j^{\text{joint}}}. \tag{56}$$

Figure 7 plots $\text{CEV}_{\text{joint}}$ against m, with a marker at the GapDIM stopping point $|\mathcal{I}|$. On the retrieval split, four of the five evaluated backbones retain roughly 89%-95% of the joint coordinate variance at the GapDIM stop. The remaining backbone retains about 72%, while still preserving retrieval and improving clustering in the main results. In every case, the stopping point occurs before the final steep increase in cumulative variance contributed by the discarded coordinates. Therefore, this confirm our intuition that GapDIM prunes the small set of coordinates that explain large variance, which is associated to the modality gap. Indeed, this supports the interpretation that GapDIM removes a small, high-gap, low-leverage tail rather than truncating the principal shared structure of the embedding space: this terminal tail contributes disproportionately to raw coordinate variance, but is not beneficial for the semantic alignment measured by retrieval and clustering.

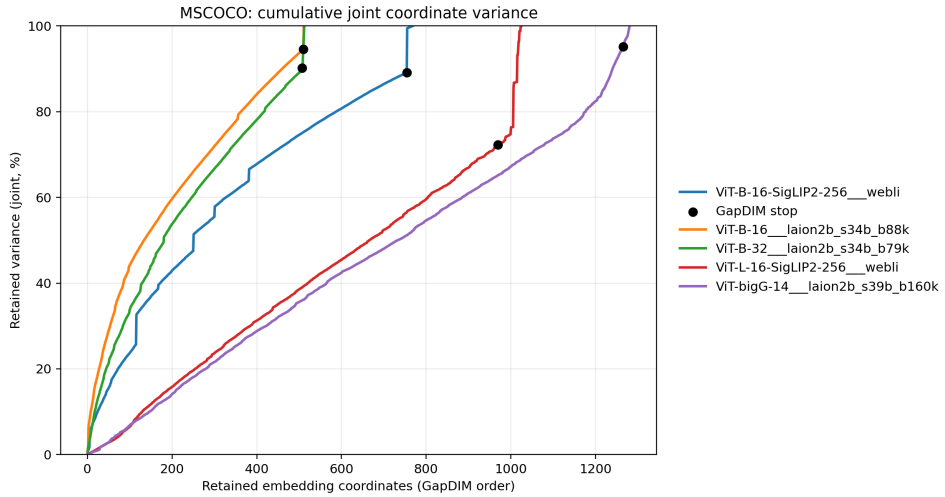

Figure 7: Cumulative joint coordinate variance on the MSCOCO retrieval split as a function of the number of GapDIM-ranked coordinates retained. Black markers denote the GapDIM stop $|\mathcal{I}|$. The stop typically occurs before the sharp terminal rise introduced by the rejected coordinates, consistent with the interpretation that the pruned tail contains noise-like gap coordinates rather than stable shared semantic structure.

E.2 Multimodal Low-Rank Controls

We run two complementary Eckart-Young controls to test whether the modality gap is itself a low-rank, easily exposable direction, or a diffuse high-dimensional offset. The first uses the row-stacked joint matrix $\mathbf{Z} = [\mathbf{X}; \mathbf{Y}]$, globally centered, so that text and image rows share a single coordinate system and the modality gap can appear directly as a low-rank direction. On MSCOCO/ViT-B/32, the leading joint principal component explains only 16.98% of the total joint energy, but captures 99.91% of the projection of $\mathbf{g}$ onto it and retains 99.96% of the gap norm. The modality gap is therefore not a diffuse offset spread across many directions: a single leading joint direction exposes almost all of it.

As a control, we build a feature-concatenated paired matrix $\mathbf{C} = [\mathbf{X}\ \mathbf{Y}] \in \mathbb{R}^{N\times 2D}$ and center the text and image blocks separately before concatenation. Let $\mathbf{C}_c = [\mathbf{X}_c\ \mathbf{Y}_c] = \mathbf{U}_C\mathbf{\Sigma}_C\mathbf{V}_C^\top$ denote the resulting block-centered matrix and its singular value decomposition. We partition its k-th right singular vector according to the two modality blocks:

$$\mathbf{v}_{C,k} = \begin{bmatrix} \mathbf{v}^X_{C,k} \\ \mathbf{v}^Y_{C,k} \end{bmatrix}, \qquad \mathbf{v}^X_{C,k}, \mathbf{v}^Y_{C,k} \in \mathbb{R}^D. \tag{57}$$

We define the text and image loading shares of component k as

$$\eta_{X,k} = \frac{\|\mathbf{v}^X_{C,k}\|_2^2}{\|\mathbf{v}^X_{C,k}\|_2^2 + \|\mathbf{v}^Y_{C,k}\|_2^2}, \qquad \eta_{Y,k} = \frac{\|\mathbf{v}^Y_{C,k}\|_2^2}{\|\mathbf{v}^X_{C,k}\|_2^2 + \|\mathbf{v}^Y_{C,k}\|_2^2}. \tag{58}$$

Since the right singular vectors have unit norm, $\eta_{X,k} + \eta_{Y,k} = 1$. This removes the centroid gap from the data matrix by construction: the gap norm changes from 0.73476 before block-centering to 9.25×10^{-16} after. For the leading singular direction, the loading shares in equation 58 are nearly balanced:

$$\eta_{X,1} = 49.49\%, \qquad \eta_{Y,1} = 50.51\%. \tag{59}$$

The projection of this direction onto the now-virtual gap axis is negligible (4.82×10^{-4}). The leading directions of the paired-concatenated control therefore involve both modality blocks and describe paired sample-to-sample variation rather than a modality-separation component.

Figure 8 contrasts the two constructions. Together, they support the central geometric assumption behind GapDIM: the removable part of the modality gap is concentrated in a small number of shared-coordinate directions, and pruning the coordinates with high gap and low participation need not discard the paired, sample-dependent semantic structure that retrieval relies on.

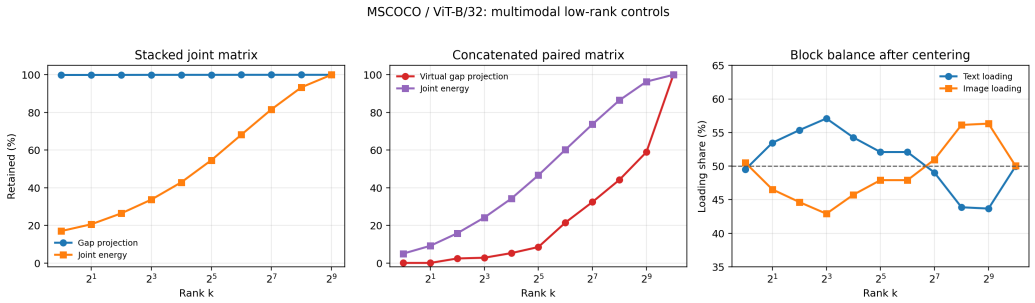

Figure 8: Multimodal low-rank controls on MSCOCO with ViT-B/32. Left: in the globally centered stacked matrix, the gap projection is almost completely captured by the first joint component, even though joint energy accumulates gradually. Middle: in the feature-concatenated paired matrix with per-block centering, the virtual gap-axis projection remains small until high rank. Right: the leading paired directions have balanced text/image loading shares, indicating that they capture paired semantic variation rather than modality separation.

Table 6: Cross-dataset threshold transfer: thresholds calibrated on Flickr30k, applied zero-shot to MSCOCO with ViT-B/32.

Method	Dim	RMG	L2M	R@1	R@10	V-measure
Pretrained	512	0.80	0.74	36.45	73.32	61.96
GapDIM (calibrated on MSCOCO)	507	0.77	0.53	36.67	73.14	71.07
GapDIM (thresholds transferred from Flickr30k)	508	0.77	0.53	36.57	73.20	62.56

F Additional Ablations

This appendix isolates the effect of the different design choices in GapDIM.

F.1 Cross-Dataset Threshold Transfer

The default method calibrates and evaluates the thresholds on the same dataset (even though in different splits). To test whether the calibration procedure is dependent by a dataset, we transfer the thresholds $(\tau^{\star}_{\text{imp}}, \tau^{\star}_{\text{gap}})$ selected on Flickr30k directly to MSCOCO, without any re-optimization: it is only the coordinate mask induced by the transferred thresholds that is applied to the MSCOCO embeddings. This is a zero-shot distribution-shift setting, since the calibration and test statistics now come from different datasets. Table 6 reports the results for ViT-B/32. Transferring the thresholds preserves almost the same coordinate mask (508 retained coordinates versus 507 under target-specific calibration) and yields retrieval and gap-reduction essentially identical to the calibrated setting, with a smaller but still positive V-measure improvement over the pretrained baseline. Target-specific calibration remains the best configuration, but this ablation demonstrates that GapDIM does not necessarily require calibration and test data to come from the same dataset/distribution to retain most of its benefit.

F.2 Cross-Model Threshold Transfer

We additionally test whether thresholds transfer across backbones with different embedding dimensions. We fix MSCOCO as the dataset and transfer the thresholds selected for ViT-B/32 ($D = 512$) directly to ViT-bigG/14 ($D = 1280$), again without re-optimization. Table 7 reports the results. Despite the substantial difference in embedding dimensionality between the two backbones, the transferred thresholds again preserve nearly all of the retrieval and clustering improvement obtained with backbone-specific calibration. Target-specific calibration for the specific dataset-model pair remains the best-performing configuration overall, but both cross-dataset and cross-model transfer show that GapDIM benefit is not tied to recalibrating on every new deployment target and that the proposed method is instead robust to distribution and backbone shifts.

Table 7: Cross-model threshold transfer: thresholds calibrated on ViT-B/32, applied to ViT-bigG/14 on MSCOCO.

Method	Dim	RMG	L2M	R@1	R@10	V-measure
Pretrained	1280	0.77	0.60	46.98	80.79	60.65
GapDIM (calibrated on ViT-bigG/14)	1265	0.75	0.50	47.55	80.89	68.56
GapDIM (thresholds transferred from ViT-B/32)	1267	0.75	0.51	47.35	80.81	67.14

Table 8: Effect of normalizing the modality centroids before computing the gap vector (equation 60), on MSCOCO with ViT-B/32.

Method	Dim	RMG	L2M	R@1	V-measure
Pretrained	512	0.804	0.736	36.45	61.96
GapDIM (default, raw centroids)	506	0.769	0.533	**36.67**	71.07
GapDIM + normalized centroids	482	**0.737**	**0.323**	32.90	**87.34**

F.3 Centroid Normalization Before Gap Computation

Because the embeddings are normalized and therefore lie on the unit hypersphere, one possible variant of the gap score projects the centroids back onto the sphere before differencing them, rather than using the raw Euclidean centroids of equation 3:

$$\widehat{\boldsymbol{\mu}}_X = \frac{\boldsymbol{\mu}_X}{\|\boldsymbol{\mu}_X\|_2}, \qquad \widehat{\boldsymbol{\mu}}_Y = \frac{\boldsymbol{\mu}_Y}{\|\boldsymbol{\mu}_Y\|_2}, \qquad \widehat{\mathbf{g}} = \widehat{\boldsymbol{\mu}}_X - \widehat{\boldsymbol{\mu}}_Y, \tag{60}$$

which replaces $\mathbf{g}$ in equation 4 while leaving the leverage computation $\ell_{M,j}$, unaffected by centroid normalization since it only uses the centered, not the raw, embedding matrices unchanged. We evaluate this variant on MSCOCO with ViT-B/32. Table 8 reports the results. Normalizing the centroids before computing the gap produces a substantially larger reduction in L2M ($0.533 \rightarrow 0.323$) and a large increase in V-measure ($71.07 \rightarrow 87.34$), consistent with a gap score that more faithfully respects the spherical geometry of the embedding space. However, R@1 decreases by nearly four points relative to the default. Selecting coordinates by their contribution to the normalized centroid displacement removes more coordinates that also carry instance-level retrieval signal, revealing a sharper retrieval-clustering trade-off than the default Euclidean gap score. This indicates that Euclidean centroid differencing, while geometrically simpler than a sphere-aware alternative, better preserves the alignment structure underlying retrieval.

F.4 Standardized GapDIM

The default GapDIM output described in Section 3.2 applies no post-slicing transformation: the retained coordinates are exactly the original pretrained coordinates, restricted to $\mathcal{I}$. We evaluate a variant, GapDIM-ST, that additionally centers and standardizes the selected embeddings after pruning:

$$\widetilde{\mathbf{X}}^{\mathrm{sel}} = \mathrm{norm}\big(\mathbf{X}^{\mathrm{sel}} - \boldsymbol{\mu}_X^{\mathrm{sel}}\big), \qquad \widetilde{\mathbf{Y}}^{\mathrm{sel}} = \mathrm{norm}\big(\mathbf{Y}^{\mathrm{sel}} - \boldsymbol{\mu}_Y^{\mathrm{sel}}\big), \tag{61}$$

where $\boldsymbol{\mu}_X^{\mathrm{sel}}, \boldsymbol{\mu}_Y^{\mathrm{sel}}$ are the centroids of the pruned embeddings and $\mathrm{norm}(\cdot)$ denotes Frobenius normalization. This post-processing step is similar in spirit to the distributional-alignment transformations used by Embedding Shift, ReAlign, and I0T (Section 2), applied here only after GapDIM coordinate selection. Table 9 compares Pretrained, GapDIM, and GapDIM-ST across five backbones and three datasets (Flickr30k, MSCOCO, and LLaVA-CC3M). GapDIM-ST reduces the modality gap much further than the default configuration and often nearly doubles V-measure relative to the already-improved GapDIM. However, this comes at a cost to retrieval in most configurations, since the post-slicing centering and standardization step alters the geometry of the retained subspace relative to the one learned during pretraining. This result directly supports our contribution: rather than always maximizing gap reduction, the proposed GapDIM configuration is deliberately tuned to preserve retrieval, and GapDIM-ST shows that a more aggressive, distribution alignment post-processing step is available whenever clustering is the primary objective.

Table 9: Pretrained, GapDIM, and Standardized GapDIM (GapDIM-ST, equation 61).

Model	Method	Flickr30k					MSCOCO					LLaVA-CC3M				
		RMG	L2M	R@1	R@10	V-Meas	RMG	L2M	R@1	R@10	V-Meas	RMG	L2M	R@1	R@10	V-Meas
ViT-B/32	Pretrained	0.80	0.69	56.86	86.56	33.11	0.80	0.74	36.45	73.32	61.96	0.80	0.71	69.74	93.85	31.18
	GapDIM (ours)	0.78	0.58	57.11	86.61	36.01	0.77	0.53	36.67	73.14	71.07	0.76	0.42	69.72	93.86	39.81
	GapDIM-ST (ours)	0.72	0.02	53.24	83.92	60.44	0.71	0.02	34.45	69.80	84.55	0.72	0.06	67.69	93.08	61.73
ViT-B/16	Pretrained	0.80	0.73	61.26	88.45	37.25	0.81	0.78	39.92	75.34	63.49	0.81	0.76	66.35	93.08	29.92
	GapDIM (ours)	0.78	0.58	61.73	88.83	41.61	0.79	0.67	40.06	75.24	65.75	0.76	0.43	66.63	93.08	35.09
	GapDIM-ST (ours)	0.71	0.02	57.13	86.59	59.31	0.71	0.02	36.47	72.21	84.64	0.72	0.05	64.92	93.08	54.16
ViT-bigG/14	Pretrained	0.77	0.58	74.27	94.34	38.56	0.77	0.60	46.99	80.80	60.65	0.78	0.58	74.22	94.51	30.10
	GapDIM (ours)	0.76	0.55	74.71	94.39	42.82	0.75	0.51	47.55	80.90	68.56	0.75	0.43	73.75	94.51	25.83
	GapDIM-ST (ours)	0.70	0.02	69.18	92.40	60.35	0.70	0.02	44.05	77.67	90.38	0.71	0.05	73.03	94.27	50.25
ViT-B/16-SigLIP2	Pretrained	0.86	1.00	73.20	94.59	27.88	0.87	1.04	51.06	83.42	57.98	0.88	1.07	75.26	93.88	29.72
	GapDIM (ours)	0.77	0.49	73.59	94.94	33.77	0.85	0.88	50.65	83.09	62.49	0.75	0.33	75.51	94.39	31.59
	GapDIM-ST (ours)	0.71	0.02	68.73	93.01	52.97	0.71	0.04	46.83	80.23	86.34	0.72	0.06	74.23	95.15	63.52
ViT-L/16-SigLIP2	Pretrained	0.85	0.97	77.33	95.45	37.88	0.86	0.99	51.02	82.13	51.85	0.87	1.03	72.79	94.75	24.46
	GapDIM (ours)	0.77	0.53	77.31	95.57	38.31	0.77	0.49	51.22	82.57	60.61	0.76	0.39	72.79	95.23	28.47
	GapDIM-ST (ours)	0.71	0.02	73.20	94.44	63.13	0.71	0.02	47.87	80.17	79.00	0.71	0.05	72.79	94.75	58.89

Table 10: Synthetic controlled experiment: 10 content coordinates +2 injected modality-translation coordinates. GapDIM removes exactly the two injected coordinates.

Setting	L2M	RMG	V-measure	X2Y R@1	Y2X R@1
Base (before pruning)	0.90	0.81	0.38	**0.87**	**0.88**
GapDIM	**0.02**	**0.26**	**0.82**	**0.87**	**0.88**

G Synthetic Controlled Experiment

The main experiments evaluate GapDIM exclusively on pretrained vision-language embeddings, where the high-gap/low-leverage pattern of Section 3 is a product of contrastive pretraining dynamics (Appendix B). To isolate whether GapDIM pruning rule itself, independent of any CLIP- or SigLIP-specific geometric artifact, correctly identifies an injected modality-translation component, we construct a fully controlled synthetic setting in which the ground-truth content and gap coordinates are known by design.

Construction. We generate 10 semantic classes. For each class, we sample 40 paired 10-dimensional content vectors per modality, drawn from a Gaussian distribution centered at a class-specific centroid shared by both modalities; this shared content component is what makes each pair semantically alignable. We then append two additional coordinates to every embedding: a constant offset added to all samples of modality X and a different constant offset added to all samples of modality Y, simulating a pure modality-translation component that carries no content information. The resulting embeddings have $10 + 2 = 12$ dimensions: 10 coordinates spanning the shared semantic content, and 2 coordinates representing the injected modality gap, mirroring the qualitative behavior of the gap/leverage distributions observed in the real embeddings of Section 3.

Results. We apply GapDIM to this 12-dimensional set with the same procedure used throughout the paper. GapDIM selects for removal exactly the two injected translation coordinates and retains all ten content coordinates, recovering the ground-truth partition without any prior knowledge of which coordinates were injected. Table 10 reports the resulting metrics. Removing the two injected coordinates reduces L2M from 0.90 to 0.02 and RMG from 0.81 to 0.26, and increases V-measure from 0.38 to 0.82, while both retrieval directions (X2Y and Y2X R@1) are exactly preserved at 0.87 and 0.88.

Why retrieval is exactly preserved. Retrieval preservation is a direct consequence of the construction. By design, the two translation coordinates contribute a multiplicative, sample-independent factor to every pairwise similarity score before pruning, since they add the same constant to every embedding of a given modality: a constant per-modality offset shifts all candidate similarities for a given query by the same amount and therefore cannot change the induced ranking. Consequently, removing these two coordinates and evaluating similarity on the remaining ten coordinates yields an identical ranking, and hence identical R@1, before and after pruning. The clustering improvement, in contrast, is substantial: removing the constant per-modality offset directly closes the artificial

separation between the two modality distributions that was driving V-measure down, exactly as predicted by the mechanism of Appendix B and the gap-leverage bound of equation 14.

Scope. This experiment supports the interpretation that GapDIM gain can arise directly from the proposed filtering criterion whenever modality-specific displacement is concentrated in coordinates that carry little content information, and multimodal alignment lives in the remaining coordinates, which Section 3 and Appendix C show is approximately the regime observed in real pretrained VLMs.

H Experimental Details and Evaluation Metrics

This appendix collects the formal definitions of the evaluation metrics used throughout the paper, together with a metric that complements RMG and L2M, and the computing resources used for all experiments.

H.1 Datasets

For each dataset, we randomly select $10{,}000$ training samples to estimate the gap and participation scores and calibrate the thresholds τ_{imp} and τ_{gap}. We then evaluate the resulting frozen coordinate mask on the complete validation set.

CIFAR-10. CIFAR-10 consists of $60{,}000$ 32×32 color images evenly distributed across 10 classes, with $50{,}000$ training images and $10{,}000$ test images Krizhevsky et al. (2009). We associate each class label with a textual description, obtaining paired visual and textual inputs for multimodal evaluation.

Flickr30k. Flickr30k Plummer et al. (2015) contains approximately $31{,}000$ images, each paired with five human-written captions. We retain the first caption for each image and extract one embedding per modality. For clustering evaluation, we assign each image a semantic pseudo-label using an ImageNet-pretrained ConvNeXt classifier Liu et al. (2022). The raw pseudo-label distribution contains 815 distinct classes for 10k validation samples. Among them, we select only the most frequent classes, following Grassucci et al. (2025).

MSCOCO. MSCOCO Lin et al. (2014) is a large-scale image-caption benchmark in which each image is annotated with multiple natural-language descriptions. We use the official 2017 splits, comprising approximately $118{,}000$ training images and $5{,}000$ validation images, each associated with up to five captions. For every image, we extract one visual embedding and one textual embedding using a single caption. We assign pseudo-labels using the same procedure adopted for Flickr30k.

LLaVA-CC3M. LLaVA Visual Instruct CC3M Pretrain 595K (LLaVA-CC3M) is a filtered subset of CC3M designed to provide more balanced concept coverage and augmented with BLIP-generated captions. We follow the same preprocessing pipeline and evaluate on a test set of $10{,}000$ samples.

H.2 Tasks

H.2.1 Cross-Modal Retrieval

We evaluate cross-modal retrieval using Recall@K, with $K \in \{1, 5, 10\}$. Given the paired, normalized embeddings $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ defined in similarity is computed as the dot product. Recall@K measures the percentage of queries whose paired sample is ranked within the top-K retrieved results:

$$\mathrm{R@}K = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\mathrm{rank}(i) \leq K] \times 100, \tag{62}$$

where $\mathrm{rank}(i)$ is the position of the ground-truth match for query i in the ranking.

H.2.2 Clustering

We assess the semantic structure of the shared embedding space by applying k-means clustering to the joint set $\mathcal{Z} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{y}_1, \ldots, \mathbf{y}_N\}$, using the ImageNet-derived pseudo-labels described

Table 11: RMG, L2M, and CosTP for all evaluated backbones. RMG and L2M are reported in their original scale; CosTP is reported in percentage (0-100). Best in bold.

Model	Method	MSCOCO			Flickr30k		
		RMG	L2M	CosTP	RMG	L2M	CosTP
ViT-B/32	Pretrained	0.80	0.74	30.50	0.80	0.69	32.27
	GapDIM (ours)	**0.77**	**0.53**	**32.11**	**0.78**	**0.58**	**32.95**
ViT-B/16	Pretrained	0.81	0.78	28.09	0.80	0.73	29.99
	GapDIM (ours)	**0.79**	**0.67**	**30.65**	**0.78**	**0.58**	**33.78**
ViT-bigG/14	Pretrained	0.77	0.60	40.21	0.77	0.58	42.91
	GapDIM (ours)	**0.75**	**0.51**	**43.69**	**0.76**	**0.55**	**43.64**
ViT-B/16-SigLIP2-256	Pretrained	0.87	1.04	14.54	0.86	1.00	16.56
	GapDIM (ours)	**0.85**	**0.88**	**20.46**	**0.77**	**0.49**	**35.00**
ViT-L/16-SigLIP2-256	Pretrained	0.85	0.99	15.06	0.85	0.97	17.91
	GapDIM (ours)	**0.77**	**0.49**	**32.31**	**0.77**	**0.53**	**36.92**

in Section 4.1. The resulting partition is scored with V-measure, the harmonic mean of homogeneity and completeness:

$$\text{V-measure} = 2 \cdot \frac{\text{homogeneity} \cdot \text{completeness}}{\text{homogeneity} + \text{completeness}} \times 100. \tag{63}$$

A higher V-measure indicates that samples sharing a pseudo-label are placed in the same cluster (homogeneity) and that each cluster is dominated by a single pseudo-label (completeness).

H.2.3 Captioning

Extended captioning task and experiment descriptions are provided in the following Appendix on Extended Downstream Evaluation.

H.3 Metrics

H.3.1 RMG

Relative Modality Gap (RMG), introduced by Schrodi et al. (2025) is defined as:

$$\text{RMG} := \frac{\frac{1}{N}\sum_{i=1}^{N} dis(\mathbf{x}_i, \mathbf{y}_i)}{\frac{1}{2N(N-1)}\left(\sum_{\substack{i,j=1 \\ i\neq j}}^{N} dis(\mathbf{x}_i, \mathbf{y}_j) + \sum_{\substack{i,j=1 \\ i\neq j}}^{N} dis(\mathbf{x}_i, \mathbf{y}_j)\right) + \frac{1}{N}\sum_{i=1}^{N} dis(\mathbf{x}_i, \mathbf{y}_i)}, \tag{64}$$

where dis is the cosine dissimilarity in $[0, 1]$. Intuitively, the numerator measures the gap where it matters, i.e., for matching pairs, and the denominator accounts for the effectively used space through an approximation using the intra-modality distances.

H.3.2 CosTP: An Alignment-Based Complement to RMG and L2M

RMG and L2M in Eqs. equation 3 quantify *global* modality separation through the centroid displacement, but they do not directly measure *instance-level* alignment. We additionally report the cosine similarity between true pairs (CosTP), following the alignment metric of Goel et al. (2022) as used by Eslami & de Melo (2025):

$$\text{CosTP} = \frac{1}{N}\sum_{i=1}^{N} \text{sim}(\mathbf{x}_i, \mathbf{y}_i), \qquad \text{CosTP} \in [-1, 1], \tag{65}$$

reported in percentage (on a 0-100 scale) in Table 11. As the equation shows, the alignment is defined as the average cosine similarity between paired image and text embedding, in the CLIP shared embedding space.

H.4 Computing Resources

GapDIM requires no pretraining or finetuning of the underlying encoders: its only computational cost is the calibration procedure (SVD estimation and threshold search) and the downstream evaluation settings above. All experiments, including the additional ablations and assessment reported in this appendix, were run on a single NVIDIA A6000 GPU with 48GB of VRAM.

H.5 Statistical Significance

We run a two-sided Wilcoxon statistical test in Tab. 1 considering GapDIM and the pretrained backbone. Across all the datasets and backbones, results show that at $p < 0.05$, the aggregate reductions in RMG and L2M and the increase in V-Measure are significant, while we preserve retrieval performance, as stated throughout the paper.

I Extended Downstream Evaluation

The main experiments evaluate GapDIM on cross-modal retrieval and joint clustering. This appendix reports two additional evaluations: a cross-modal generation task that directly probes whether the coordinates GapDIM removes impact cross-modal generation, and an intra-modality evaluation that verifies GapDIM does not improve the joint space at the expense of the internal organization of either modality alone.

I.1 Image-Free Zero-Shot Captioning

Fragmented latent spaces that scatter semantic concepts across separate modality clusters are expected to be especially harmful in tasks where a decoder trained on one modality embeddings must operate at test time on the other modality embeddings, since the decoder never observes any cross-modal correspondence during training. We use image-free zero-shot image captioning as such a task, following the cross-modal transfer setting of Zhang et al. (2024).

Setup. The decoder never receives images during training. It is trained to reconstruct an MSCOCO caption $\mathbf{y}$ from its frozen CLIP text embedding:

$$\mathbf{y} \;\rightarrow\; g_\phi(\mathbf{y}) \;\rightarrow\; f_\psi(g_\phi(\mathbf{y})), \tag{66}$$

where g_ϕ is the frozen CLIP text encoder and f_ψ is the trainable decoder. At test time, the text embedding is replaced by the frozen CLIP embedding of the corresponding image:

$$x \;\rightarrow\; f_\theta(\mathbf{x}) \;\rightarrow\; f_\psi(f_\theta(\mathbf{x})), \tag{67}$$

with f_θ the frozen CLIP image encoder. The decoder f_ψ is a ClipCap-style architecture (Mokady et al., 2021): a one-hidden-layer MLP that maps the frozen CLIP embedding to a prefix of length 10 in the input space of a GPT-2-base language model (Radford et al., 2019). The MLP and GPT-2 are trainable; the CLIP encoders g_ϕ, f_θ remain frozen throughout, both when using the pretrained embeddings and when using the GapDIM-filtered embeddings. We run the experiment on MSCOCO with ViT-B/32 across three random seeds and report the mean.

Results. Table 2 (main text, Section 4) reports BLEU-1, BLEU-4, CIDEr, and SPICE for the pretrained embeddings and for the GapDIM-filtered embeddings. GapDIM improves all the metrics with respect to the pretrained baseline. Because the decoder never observes any image-conditioned supervision, this result is not explained by improved retrieval ranking: it directly indicates that the coordinates GapDIM prunes act as a causal bottleneck for the cross-modal transfer of the decoder trained purely on text embeddings, and that removing them consistently improves a cross-modal generation task beyond retrieval and clustering.

I.2 Intra-Modality Preservation

GapDIM coordinate mask $\mathcal{I}$ is shared across modalities (equation 10), so it is natural to ask whether the joint improvements reported in Section 4 come at the cost of degrading the representation of

Table 12: Intra-modality linear probing on MSCOCO with ViT-B/32 (517-class pseudo-labels, top-1 accuracy).

Modality	Pretrained	GapDIM
Image-only	60.59%	**61.21**%
Text-only	52.27%	**52.33**%

Table 13: Intra-modality unimodal k-means clustering ($k = 10$) on MSCOCO with ViT-B/32 (V-measure).

Modality	Pretrained	GapDIM
Image-only	85.79%	**87.75**%
Text-only	85.68%	**88.65**%

either modality considered in isolation. We evaluate image-only and text-only representations independently on MSCOCO with ViT-B/32, comparing the pretrained embeddings with the GapDIM-filtered embeddings.

Linear probing. We freeze the encoders and fit a one-vs-rest logistic-regression head independently on the image embeddings and on the text embeddings of the MSCOCO training split, using the 517-class pseudo-label set described in Section 4.1, and report top-1 accuracy on the validation split. Table 12 reports the results.

Unimodal clustering. We run k-means with $k = 10$ separately on the image embeddings and on the text embeddings (rather than on the row-stacked joint set $\mathcal{Z}$ used in the main clustering evaluation), using the same pseudo-labels, and report V-measure. Table 13 reports the results.

Results. GapDIM slightly improves both linear-probe accuracy and unimodal clustering quality in every modality/task combination, rather than trading intra-modality quality for joint-space improvement. This is consistent with the coordinates removed by GapDIM being dominated by a cross-modal-translation component with low leverage in each modality own dominant subspace (Section 3): removing them has a small, uniformly non-negative effect on tasks that look at only one modality at a time, showing that GapDIM not only does not hurt unimodal distribution organization but slightly improves it.

J LIMITATIONS

Although GapDIM shows consistent behavior across all backbones, datasets, and assessment evaluated in this paper, we discuss several limitations.

Dependence on embedding dimensionality. GapDIM benefits from having a sufficiently large embedding dimension D to prune, since the method removes a small fraction of coordinates rather than restructuring the retained ones. It may therefore be less effective in backbones with much smaller embedding sizes, where fewer coordinates are available to isolate the high-gap, low-leverage pattern. This is partly mitigated by the trend in multimodal learning toward larger embedding dimensions, as reflected in the range of backbones evaluated in this paper ($D = 512$ to $D = 1280$).

Scope of the theoretical mechanism. The high-gap/low-leverage pattern that motivates GapDIM is grounded in an approximate-orthogonality mechanism for contrastive pretraining (Appendix B) and validated empirically across the CLIP and SigLIP families studied in this paper (Appendix C), and directly reproduced in a fully controlled synthetic construction (Appendix G). This does not constitute a proof that the pattern holds for every VLM, or that it extends to architectures or training objectives not evaluated here. We view GapDIM selection criterion as a measurable, empirically validated regularity of current contrastive VLMs, rather than as a property guaranteed to hold by construction for arbitrary fragmented multimodal representations.