MODALITY GAP-DRIVEN SUBSPACE ALIGNMENT TRAINING PARADIGM FOR MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the success of multimodal contrastive learning in aligning visual and linguistic representations, a persistent geometric anomaly, the Modality Gap, remains: embeddings of distinct modalities expressing identical semantics occupy systematically offset regions. Prior approaches to bridge this gap are limited by oversimplified isotropic assumptions, hindering their application in large-scale scenarios. In this paper, we address these limitations by characterizing the geometric shape of the modality gap and leveraging it for efficient model scaling. We reveal the structured geometric characteristics of the Modality Gap: it not only contains a global centroid shift, but also exhibits significant anisotropic residuals, while the two modalities still share similar dominant geometric structures. Guided by this precise modeling, we introduce ReAlign, a training-free modality alignment strategy. Utilizing statistics from massive unpaired data, ReAlign aligns text representation into the image representation distribution via Anchor, Trace, and Centroid Alignment, thereby explicitly rectifying geometric misalignment. Building on ReAlign, we propose ReVision, a scalable pre-training paradigm for Multimodal Large Language Models (MLLMs). ReVision integrates ReAlign into the pretraining stage, enabling the model to learn the distribution of visual representations from unpaired text before visual instruction tuning, without large-scale, high-quality image-text pairs. Our framework demonstrates that statistically aligned unpaired data can effectively substitute for expensive image-text pairs, offering a robust path for the efficient scaling of MLLMs.

1 INTRODUCTION

Multimodal contrastive learning (Huang et al., 2024; Zhai et al., 2023; Radford et al., 2021; Sun et al., 2023; Yaras et al., 2024) has established itself as the standard paradigm for aligning visual and text representations. However, despite extensive training on massive image-text pairs, a persistent empirical observation remains:

Phenomenon. Modality Gap

For two modalities that express the same underlying semantics, their representations typically do not coincide. Instead, the two modalities inherently occupy distinct, systematically offset regions of the joint representation space. This phenomenon is called the ***Modality Gap*** *(Liang et al., 2022; Zhang et al., 2024).*

While this gap hinders direct cross-modal interchangeability, it also suggests that, if the geometric misalignment can be effectively bridged, abundant text data could serve as a substitute (Yu et al., 2025) for expensive image-text data in Multimodal Large Language Models (MLLMs) training.

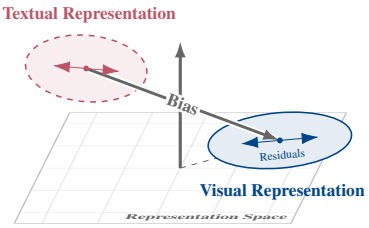


Figure 1: Modality Gap Phenomenon.

Previous research has largely advanced along two main directions, yet both face limitations:

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

① **Geometric Correction:** These approaches attempt to post-hoc correct this gap via explicit geometric projections (Zhang et al., 2024; Yi et al., 2025). However, most of these works are limited to single, small-scale tasks (Tewel et al., 2022; Li et al., 2023a; Liu et al., 2024; Su et al., 2022; Nukrai et al., 2022) such as Image Captioning, failing to unlock the true potential of modality gap theory for model scaling. Furthermore, they typically rely on isotropic noise assumptions, failing to account for the complex anisotropic structures in high-dimensional spaces. ② **Text-only Large-scale Training:** These methods (Yu et al., 2025) leverage modality gap to synthesize pseudo-visual supervision signals from pure text. While promising, existing methods suffer from substantial performance degradation on fine-grained visual tasks, exposing a pronounced distributional gap between synthetic text representations and real-world image data.

These limitations point to a fundamental mismatch: existing assumptions are overly simplified, resulting in a lack of precise modeling of the modality gap's geometric shape, which in turn hinders its application in large-scale training. This mismatch naturally motivates two core research questions:

① **On Shape:** Can we move beyond simple mean assumptions to precisely characterize the intrinsic geometric shape of the modality gap within a stable reference frame?
② **On Scale:** Can we leverage this precise shape modeling to design a scalable training paradigm that substitutes expensive paired data with massive, easily accessible unpaired data, thereby achieving efficient MLLM scaling?

To address the first question regarding *shape*, we conduct a geometric analysis of the Modality Gap across multiple pretrained vision–language encoders. We find that the Modality Gap cannot be explained solely by a global centroid shift; after removing the centroid displacement, the remaining discrepancy still exhibits significant anisotropic structure rather than simple isotropic noise. More importantly, despite the substantial cross-modal offset, image and text representations retain highly consistent spectral decay and substantially overlapping dominant subspaces. We further observe that spherical normalization can reintroduce centroid misalignment, a phenomenon we refer to as *Phantom Drift*. These findings suggest that the Modality Gap is fundamentally a structured geometric mismatch with clear statistical regularities, and that effective alignment should correct first-order displacement while preserving the existing dominant geometry shared by the two modalities.

Building on this precise geometric modeling, we further address the second question regarding scale. Our approach highlights a perspective different from prior work: while acquiring high-quality paired image-text data (He et al., 2024; Chen et al., 2024b; Li et al., 2024) is expensive, obtaining massive amounts of unpaired single-modality data is extremely easy. We believe that leveraging large-scale unpaired data is sufficient to estimate the statistical mismatches between modalities without relying on paired samples. Guided by this insight, we introduce two core contributions:

① **ReAlign.** Building on the geometric analysis of Modality Gap, we introduce ReAlign, a training-free pre-alignment strategy that maps text representations into the image representation distribution using statistics derived from large-scale unpaired data. ReAlign operates through a three-stage procedure. First, Anchor Alignment matches first-order statistics (means). Second, Trace Alignment matches the global variance scale. Finally, Centroid Alignment explicitly rectifies the geometric drift induced by spherical projection. Together, these stages achieve precise cross-modal alignment at the statistical level using only linear transformations and normalization, without requiring any training.

② **ReVision.** We introduce ReVision, a scalable pre-training paradigm for MLLMs. Building on the geometry-preserving ReAlign strategy, ReVision maps large-scale unpaired text representations into the image representation distribution to generate pseudo-visual representations. The adapter is then trained while keeping the LLM backbone frozen, enabling low-cost semantic injection from large-scale text corpora and allowing the model to absorb broader world knowledge without relying on expensive, high-quality image–text pairs. Since ReAlign preserves the rich semantics of text while preserving the shared anisotropic geometry and matching the global second-order scale of image representations, the resulting pseudo-visual representations can serve as effective visual surrogate signals for MLLM pretraining.

Overall, we characterize the Modality Gap as a structured geometric mismatch. Based on this analysis, we introduce a training-free alignment strategy that preserves shared geometry and enables scalable use of unpaired data for cost-efficient MLLM training.

2 Task Setting

We consider a contrastively pretrained vision–language model, where both the image encoder E_{img} and text encoder E_{text} are kept frozen. Given two sets of unpaired image samples $\{x_i\}_{i=1}^{N_x}$ and text samples $\{y_j\}_{j=1}^{N_y}$, we obtain the image representations $\{e_{x_i}\}$ and text representations $\{e_{y_j}\}$ through E_{img} and E_{text}, respectively. We use the distributional statistics of these two sets of representations to construct a modality transformation $\mathcal{S}_{y\to x}$, which transfers text representations into the image representation space without relying on any image–text pairing:

$$\tilde{e}_x = \mathcal{S}_{y\to x}(e_y).$$

Once this transformation is established, arbitrary text can be encoded by E_{text} and further mapped to the corresponding pseudo-visual representation $\tilde{e}_x$ for subsequent MLLM pretraining. Therefore, large-scale unpaired text corpora can be used to replace expensive paired image–text data.

3 Modality Gap Analysis

To analyze the geometric structure of the Modality Gap, we select six contrastively pretrained vision–language encoders with different architectures and pretraining data, including LLM2CLIP (Huang et al., 2024), MetaCLIP 2 (Chuang et al., 2026), SigLIP2 (Tschannen et al., 2025), as well as OpenCLIP-80M (Cherti et al., 2023), OpenCLIP-400M (Cherti et al., 2023), and OpenCLIP-2B (Cherti et al., 2023) at different scales. For semantically paired image and text representations $e_x, e_y \in \mathbb{R}^d$, we conduct our analysis around three progressive questions in Subsection 3.1–3.3:

3.1 Insufficiency of Centroid Shift

According to prior works, for semantically paired image and text representations e_x and e_y, the Modality Gap $\Delta = e_x - e_y$ is modeled as $\Delta = c_\perp + \epsilon$ (Zhang et al., 2024), where $c_\perp$ denotes a stable global modality offset, while ϵ is regarded as random noise. Such a formulation implicitly assumes that the main structure of the Modality Gap can be explained by a global offset vector.

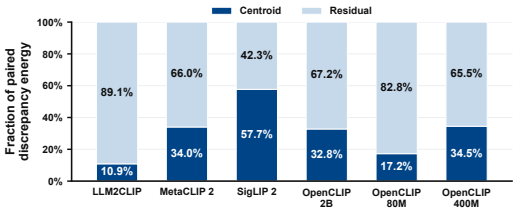


Figure 2: Decomposition of the Modality Gap into centroid and residual components across six vision–language encoders. The residual term accounts for the majority of the paired discrepancy energy in five of the six models.

However, whether this assumption generally holds across different vision–language encoders has not been directly verified. In particular, if the centroid shift cannot explain most of the cross-modal discrepancy, then simply translating the global centers of the two modalities may be insufficient for effective alignment. We therefore first examine how much of the Modality Gap can be explained by the global centroid shift. Specifically, we decompose the paired discrepancy as:

$$\Delta = \mathbb{E}[\Delta] + r,$$

where $\mathbb{E}[r] = 0$. Since the cross term vanishes in expectation, the total paired discrepancy energy can be exactly decomposed as $\|\mathbb{E}[\Delta]\|_2^2 + \mathbb{E}\|r\|_2^2$. Based on this decomposition, we compute the proportions of the centroid term and the remaining term in the total discrepancy energy to examine whether the Modality Gap is primarily composed of a global centroid shift. As shown in Fig.2, the centroid term typically explains only a small portion of the Modality Gap: in five of the six models, the remaining discrepancy energy exceeds 50%, with LLM2CLIP and OpenCLIP-80M reaching 89.1% and 82.8%. Even when considering all six models, approximately 68.8% of the discrepancy energy on average cannot be explained by the global centroid shift.

Conclusion 1.

The Modality Gap cannot be primarily explained by a global centroid shift, as most of the cross-modal discrepancy remains in the residual after centroid correction.

This further raises the next question: If the centroid shift is insufficient to explain the overall discrepancy, can the remaining residual be approximately regarded as isotropic noise?

3.2 ANISOTROPY RESIDUAL

After removing the centroid shift, we further examine whether the remaining discrepancy can be approximately regarded as isotropic noise. Let the centered cross-modal discrepancy be $r = (e_x - \mu_x) - (e_y - \mu_y)$. If r were primarily composed of isotropic random noise, its energy should be distributed approximately uniformly across all directions, and the corresponding covariance matrix should satisfy $\Sigma_r = \mathrm{Cov}(r) \approx \sigma^2 I$. In this case, the eigenvalues of Σ_r should be approximately equal. Therefore, we first compare the largest eigenvalue with the mean eigenvalue, and use:

$$A_r = \frac{\lambda_{\max}(\Sigma_r)}{\mathrm{tr}(\Sigma_r)/d}$$

to measure the concentration of the remaining discrepancy energy along the dominant direction. For an ideally isotropic distribution, all eigenvalues are equal and thus $A_r = 1$; a larger A_r indicates that the energy is more concentrated in a small number of directions.

As shown in Fig.3, the A_r values of the six models are 28.6, 74.8, 78.7 26.8, 18.6, and 26.5, respectively, all far above the isotropic baseline. This means that the variance along the strongest direction of the remaining discrepancy is approximately 19 to 79 times larger than the average directional variance, indicating that the discrepancy remaining after centroid correction cannot be approximated as isotropic noise and instead exhibits clear directional structure.

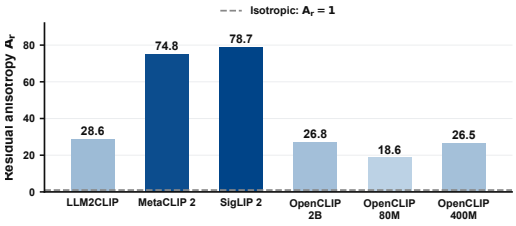


Figure 3: Residual anisotropy across six vision–language encoders after centroid correction. All models exhibit A_r values far above the isotropic baseline $A_r = 1$.

Examining only the largest eigenvalue is still insufficient to characterize the overall structure. We therefore further analyze the full eigenspectrum of Σ_r. To eliminate the effect of different representation dimensionalities across models, we use the normalized spectrum:

$$a_j = d\lambda_j(\Sigma_r)/\,\mathrm{tr}(\Sigma_r).$$

If the remaining discrepancy were isotropic, all eigenvalues would be equal, and $a_j = 1$ would hold for any j.

Fig.4 shows that the spectra of all six models deviate substantially from this baseline: the leading eigenvalues are strongly amplified, followed by a continuous decay as the eigenvalue rank increases, eventually falling below the isotropic baseline in the spectral tail. This indicates that the energy of the remaining discrepancy is not uniformly distributed throughout the representation space, but is systematically concentrated along a limited set of dominant directions.

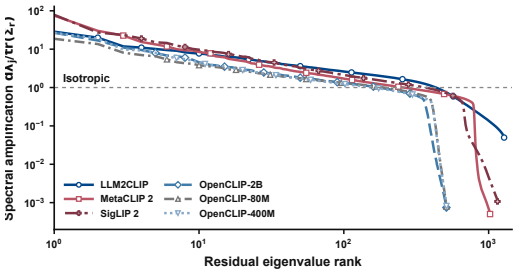
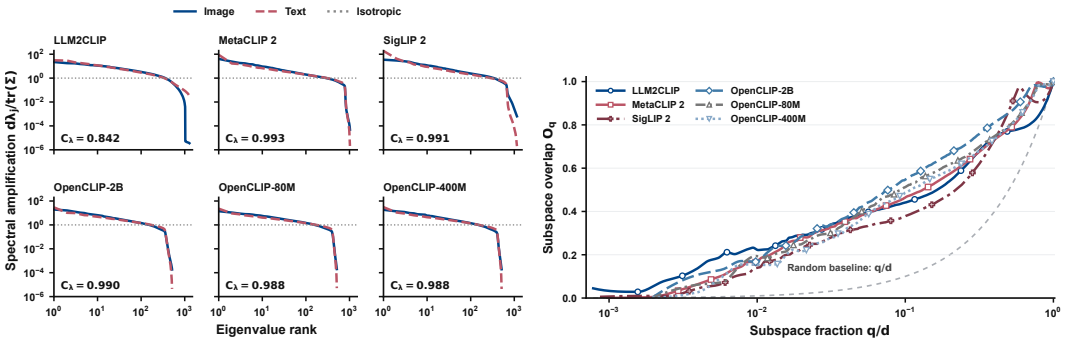


Figure 4: Spectral amplification of the residual cross-modal discrepancy across six vision–language encoders. All models deviate substantially from the isotropic baseline $a_j = 1$.

Conclusion 2.

The discrepancy remaining after centroid correction is not isotropic random noise, but exhibits significant anisotropic structure.

If the dominant geometries of the two modalities are unrelated, then cross-modal alignment would still require paired data to learn their correspondence. This further raises a new question: in the presence of significant anisotropic structure, do image and text still share similar dominant representation geometry?



(a) (b)

Figure 5: (a) Trace-normalized eigenspectra of image and text representations across six vision–language encoders. The two modalities exhibit highly consistent spectral decay. (b) Overlap O_q between the top-q image and text eigenspaces as a function of the subspace fraction q/d. The observed overlap remains well above the random-subspace baseline q/d, particularly for low and intermediate subspace fractions. Together, these results show that image and text preserve similar spectral structure and substantially overlapping dominant subspaces despite the Modality Gap.

3.3 Consistency of Dominant Geometry

We compare the second-order geometry of image and text representations from two complementary perspectives: spectral structure and dominant subspaces.

We first normalize the eigenspectra of the centered covariance matrices $\Sigma_x = \mathrm{Cov}(e_x - \mu_x)$ and $\Sigma_y = \mathrm{Cov}(e_y - \mu_y)$ by their total variance:

$$p_j^x = \frac{\lambda_j(\Sigma_x)}{\mathrm{tr}(\Sigma_x)}, \qquad p_j^y = \frac{\lambda_j(\Sigma_y)}{\mathrm{tr}(\Sigma_y)}.$$

This normalization removes the overall scale difference and focuses the comparison on the relative allocation of variance across principal directions. We further use $C_\lambda = \mathrm{corr}(\log p^x, \log p^y)$ to measure spectral consistency.

As shown in Fig. 5(a), the C_λ values of MetaCLIP 2, OpenCLIP-2B, OpenCLIP-80M, OpenCLIP-400M, and SigLIP 2 are 0.993, 0.990, 0.988, 0.988, and 0.991, respectively, all close to 0.99; even LLM2CLIP, which has the lowest correlation, reaches 0.842. The full eigenspectra further show that image and text exhibit highly similar spectral decay trends across most of the dominant region.

However, similar eigenspectra do not guarantee that the two modalities share the same principal directions. We therefore further compare their dominant eigenspaces. Let $U_x^{(q)}$ and $U_y^{(q)}$ contain the top-q eigenvectors of Σ_x and Σ_y, respectively, and define

$$O_q = \frac{1}{q}\left\|\left(U_x^{(q)}\right)^\top U_y^{(q)}\right\|_F^2.$$

For two independent random q-dimensional subspaces, the expected overlap is approximately q/d. As shown in Fig. 5(b), the actual O_q values of all six models are significantly higher than the random baseline, particularly for smaller q/d, indicating significant consistency between the dominant directions of image and text representations.

Conclusion 3.

The two modalities retain similar dominant geometric structures despite the presence of the Modality Gap.

Since image and text already exhibit similar dominant geometry in the representation space, does spherical normalization preserve this alignment?

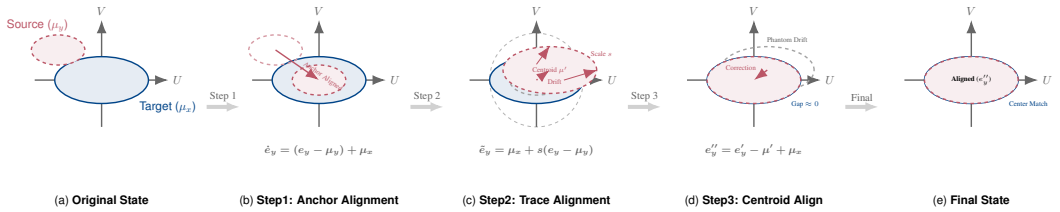


Figure 6: The ReAlign pipeline. ReAlign sequentially performs Centroid Alignment, Trace Alignment, and Centroid Alignment to reduce cross-modal statistical mismatch while preserving the centered second-order geometry.

4 REALIGN: TRAINING-FREE MODALITY ALIGNMENT

Sec.3. shows that the Modality Gap cannot be explained solely by centroid displacement: the remaining discrepancy exhibits significant anisotropy, while the two modalities retain highly similar normalized spectra and substantially overlapping dominant subspaces. This observation suggests that effective alignment should not reconstruct the second-order geometry from scratch. Instead, it should preserve the anisotropic structure already shared across modalities while correcting the statistical mismatches that remain. Based on this principle, we propose **ReAlign**, a training-free and anisotropy-preserving modality alignment strategy, as illustrated in Fig. 6. ReAlign does not require paired image-text data or sample-level correspondence. It only uses two independently sampled sets of image and text representations to estimate modality-level statistics. ReAlign sequentially corrects first-order displacement, calibrates the global second-order scale without altering the normalized anisotropic geometry, and compensates for the centroid drift introduced by spherical normalization.

4.1 STEP 1: CENTROID ALIGNMENT

Motivation. Subsec.3.1 shows that a stable global centroid displacement exists between the two modalities. Although this displacement is insufficient to explain the entire Modality Gap, it constitutes a directly estimable first-order discrepancy. ReAlign therefore first removes this centroid displacement through translation.

Operation. Let $e_y, e_x \in \mathbb{R}^d$ denote the unit-normalized embeddings of the source modality y and target modality x, with means $\mu_y = E[e_y]$ and $\mu_x = E[e_x]$. We center the source representation and translate it to the target centroid:

$$\dot{e}_y = (e_y - \mu_y) + \mu_x.$$

This gives $E[\dot{e}_y] = \mu_x$, while translation leaves the centered covariance unchanged and therefore preserves the original second-order geometry.

4.2 STEP 2: TRACE ALIGNMENT

Motivation. After centroid translation, the remaining cross-modal discrepancy is strongly anisotropic rather than isotropic noise. More importantly, Subsec.3.3 shows that image and text representations already exhibit highly consistent trace-normalized eigenspectra and substantially overlapping dominant subspaces. Therefore, the goal of ReAlign is not to reconstruct the target-modality covariance or enforce full covariance matching. Such aggressive covariance transformations may unnecessarily distort the semantic geometry already shared across the two modalities. Instead, we preserve the relative anisotropic variance structure of the source modality and only calibrate its overall second-order scale to that of the target modality.

Operation. Let $\Sigma_y = \mathrm{Cov}(e_y - \mu_y)$ and $\Sigma_x = \mathrm{Cov}(e_x - \mu_x)$. We define the total variances of the two modalities as $T_y = \mathrm{tr}(\Sigma_y)$, and $T_x = \mathrm{tr}(\Sigma_x)$. We then apply the global trace-matching scale:

$$s = \sqrt{\frac{T_x}{T_y + \epsilon}}, \quad \tilde{e}_y = \mu_x + s(e_y - \mu_y),$$

where ϵ is a small constant for numerical stability. After scaling, the centered covariance becomes $\tilde{\Sigma}_y = s^2 \Sigma_y$. Therefore, Trace Alignment changes the variance scale while leaving the covariance

eigenvectors unchanged. Meanwhile, the trace-normalized eigenspectrum is exactly preserved:

$$\frac{\lambda_j(\tilde{\Sigma}_y)}{\mathrm{tr}(\tilde{\Sigma}_y)} = \frac{\lambda_j(\Sigma_y)}{\mathrm{tr}(\Sigma_y)}.$$

Thus, Trace Alignment can be more precisely characterized as anisotropy-preserving scale calibration, whose core objective is to calibrate the overall second-order scale while preserving the original covariance geometry. This point is important: ReAlign directly leverages the geometric structure already shared across the two modalities and reduces their statistical discrepancy through global scale calibration, thereby avoiding excessive reshaping of the source-modality covariance structure. Appendix D further compares this design with stronger covariance-matching transformations.

The above translation and scaling generally violate the unit-norm constraint of the original vision–language representation space. We therefore project the calibrated representation back onto the unit hypersphere $e'_y = \frac{\tilde{e}_y}{\|\tilde{e}_y\|}$.

4.3 Step 3: Centroid Alignment

Motivation. Spherical normalization does not preserve the centroid alignment established in Euclidean space. Because normalization is nonlinear, projecting $\tilde{e}_y$ onto the unit hypersphere reintroduces a centroid displacement, which we refer to as **Phantom Drift**. Therefore, ReAlign performs an additional centroid correction after spherical projection.

Operation. Let $\mu'_y = \mathbb{E}[e'_y]$ denote the centroid after spherical projection. The Phantom Drift relative to the target centroid μ_x is $\Delta_y^{\pi} = \mu'_y - \mu_x$. We remove this drift by translating the projected representation back to the target centroid:

$$e''_y = e'_y - \mu'_y + \mu_x.$$

This correction exactly restores the target centroid, i.e., $\mathbb{E}[e''_y] = \mu_x$, while keeping the centered covariance unchanged, i.e., $\mathrm{Cov}(e''_y) = \mathrm{Cov}(e'_y)$. Although the final translation does not strictly preserve the unit-norm constraint, it only introduces a shared first-order correction and does not alter the centered second-order geometry. Additional theoretical analysis of ReAlign is provided in Appendix A. The resulting norm deviation is analyzed in Appendix B.6.

5 ReVision: A Scalable MLLM Pre-training Paradigm

We introduce **ReVision**, a pre-training paradigm for MLLMs. Building on the geometry-preserving ReAlign strategy, ReVision synthesizes pseudo-modality features from unpaired text corpora. This design enables low-cost semantic injection, allowing the model to absorb extensive world knowledge during pretraining without relying on expensive, high-quality paired data.

Modality Substitution Pretraining. Let E_{img} and E_{text} denote the frozen image and text encoders. Freed from the reliance on paired visual-text data, we leverage E_{text} to encode large-scale unpaired text corpora. Unlike traditional multimodal datasets limited by data scarcity, this strategy allows training to scale to raw text resources. As a result, the model receives dense semantic supervision and can absorb broader world knowledge during pretraining than standard approaches.

We define a training-free modality substitution operator $S_{y\to x}$ based on Section 4. Given a text sample y, this operator maps its embedding into the image space distribution:

$$\tilde{e}_x = S_{y\to x}(E_{\text{text}}(y)).$$

Here, $\tilde{e}_x$ serves as a pseudo-visual embedding. Because ReAlign is geometry-preserving, it preserves the rich semantics of y while calibrating the pseudo-visual representation toward the global geometry of real image representations. We train the adapter ϕ (with LLM θ frozen) to reconstruct the text conditioned on the pseudo-visual embedding $\tilde{e}_x$:

$$\mathcal{L}(\phi) = -\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, T_\phi(\tilde{e}_x)).$$

6 EXPERIMENTS

In this section, we systematically investigate the effectiveness of ReVision under the proposed geometry-preserving framework. Our experiments are designed to address three core research questions (RQs). To ensure a fair comparison, all methods employ the same model architecture. Detailed training and evaluation settings, together with the controlled pretraining-data comparison, are provided in Appendix E.

RQ1. How effectively does ReAlign correct cross-modal statistical mismatch while preserving representation geometry? We first evaluate the alignment effectiveness of ReAlign by measuring the Euclidean distance between the image and text centroids. We use LLM2CLIP-OpenAI-L-14-336 as the encoder and perform alignment on 100K samples from the Bunny-Pretrain and DenseFusion datasets. As illustrated in Figure 7, the original representations exhibit substantial centroid gaps of 0.3918 and 0.4276 on Bunny and DenseFusion, respectively. C^3 substantially reduces the first-order mismatch, reaching approximately 0.0023 on both datasets. ReAlign further reduces the residual centroid gap to 2.64×10^{-4} on Bunny and 1.39×10^{-4} on DenseFusion, demonstrating more accurate first-order statistical alignment across different data distributions. Centroid distance, however, characterizes only first-order alignment and does not by itself measure whether the underlying representation geometry is preserved. We therefore complement this analysis with distribution-level geometric evaluations in Appendix B. Across six vision-language encoders, ReAlign largely preserves the trace-normalized spectral structure and dominant-subspace overlap already present in the original text representations, while producing pairwise angular distributions substantially closer to those of the image representations than C^3. These results indicate that ReAlign can correct the dominant cross-modal statistical mismatch while avoiding unnecessary distortion of the structured geometry already shared between the two modalities. We further evaluate the calibration efficiency, numerical stability, computational scalability, and domain sensitivity of ReAlign in Appendix C.

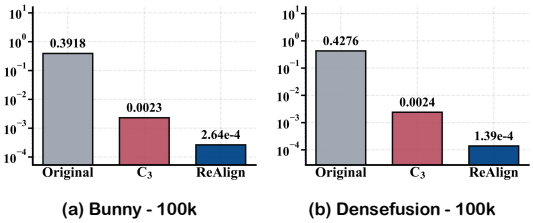


Figure 7: **Cross-modal centroid alignment on Bunny-Pretrain and DenseFusion.** C^3 substantially reduces the original centroid mismatch, while ReAlign further lowers the residual gap to the 10^{-4} scale on both datasets.

RQ2. How effective is ReAlign in large-scale MLLM training scenarios? For fair comparison, all baselines adopt the same two-stage training paradigm as ReVision: the pretraining stage uses 1M text samples from Bunny-pretrain, followed by SFT stage using InternVL-Chat-V1-2-SFT(Chen et al., 2024d). We evaluate ReVision under three vision-language backbone combinations: LLM2CLIP-Qwen3, using LLM2CLIP-Openai-L-14-336 (Huang et al., 2024) as the vision encoder and Qwen3-8B (Yang et al., 2025) as the language model; SigLIP2-Qwen3, using SigLIP2-So400m-patch14-384 (Tschannen et al., 2025) as the vision encoder and Qwen3-8B as the language model; and LLM2CLIP-Llama3, using LLM2CLIP-Openai-L-14-336 as the vision encoder and Llama-3-8B-Instruct as the language model. We compare ReVision against ① **Blind**: without providing image inputs, we use Qwen3-235B-A22B-Instruct (Yang et al., 2025) to answer text-only questions; ② **No Align**: directly using raw text representations; ③ **C^3 Align**: performing centroid alignment and injecting isotropic Gaussian noise. As shown in Table 1, ReVision achieves the highest average performance under all three backbone settings, reaching 54.39, 53.58, and 50.16 with LLM2CLIP-Qwen3, SigLIP2-Qwen3, and LLM2CLIP-Llama3, respectively, all outperforming their corresponding C^3 Align baselines. The advantage of ReVision is particularly evident on reasoning-intensive tasks, suggesting that the isotropic noise introduced by C^3 may induce a whitening effect that disrupts the fine-grained semantic hierarchy required for complex reasoning, whereas ReAlign better preserves the spectral decay and structural information of the features. Meanwhile, ReVision also shows consistent advantages on hallucination-related metrics; for example, under the LLM2CLIP-Qwen3 setting, it achieves 83.47 on CRPE and 49.82 on HallBench. More experiments are provided in Appendix F.

RQ3. Can scaling up text-only pretraining surpass the performance of paired image-text pretraining? Finally, we investigate whether ReVision can challenge the traditional paradigm that relies on paired image-text data. To ensure fairness, we sampled a subset of 417k examples from the

Table 1: Performance comparison of different geometric alignment strategies.

Method	General				Reasoning				Hallucination			Avg. ↑
	MME	MMStar	SQA	RealWorldQA	MMMU	MMMU-P	VisuLogic	LogicVista	CRPE	POPE	HallBench	
Blind	3.37	8.80	6.17	5.36	19.60	12.44	0.30	1.56	12.90	0.60	15.25	7.85
LLM2CLIP - Qwen3												
No Align	77.36	39.07	75.26	48.19	38.08	33.41	30.22	25.57	82.43	71.47	45.29	51.49
C^3 Align	80.03	38.57	75.93	47.66	39.88	35.24	31.19	24.31	81.62	72.36	46.37	52.11
ReVision (ours)	**83.69**	**39.94**	**77.08**	**52.73**	**41.27**	**36.71**	**33.68**	**27.86**	**83.47**	**72.52**	**49.82**	**54.39**
SigLIP2 - Qwen3												
No Align	76.24	39.51	75.14	47.06	36.73	32.16	29.94	24.87	81.69	71.93	44.46	50.88
C^3 Align	79.08	38.66	75.39	46.77	38.34	33.81	31.02	23.49	80.91	72.54	45.73	51.43
ReVision (ours)	**82.57**	**40.13**	**76.49**	**51.27**	**39.54**	**35.08**	**33.16**	**26.64**	**82.88**	**72.79**	**48.84**	**53.58**
LLM2CLIP - Llama3												
No Align	73.63	35.73	75.23	43.53	28.82	25.38	24.40	21.03	80.82	71.59	42.38	47.50
C^3 Align	76.16	34.60	75.52	43.14	30.69	27.20	25.50	19.91	79.99	72.43	43.53	48.06
ReVision (ours)	**79.65**	**36.13**	**76.71**	**47.97**	**31.51**	**28.39**	**27.70**	**22.82**	**81.78**	**72.53**	**46.58**	**50.16**

Table 2: Performance comparison between paired image-text pretraining and ReVision scaling.

Method	Text-only	General				Reasoning				Hallucination			Avg. ↑	Stage-1 Data
		MME	MMStar	SQA	RealWorldQA	MMMU	MMMU-P	VisuLogic	LogicVista	CRPE	POPE	HallBench		
Paired image-text	✗	73.59	35.40	76.01	44.18	34.80	27.70	25.90	23.27	**80.87**	69.13	47.11	48.91	1M image-text
Unicorn Yu et al. (2025)	✓	60.24	35.13	68.81	42.35	**36.87**	**34.05**	**26.80**	**29.53**	42.32	64.21	43.01	43.94	1M text
ReVision-1M (ours)	✓	72.20	34.33	75.84	43.72	30.22	27.64	25.70	21.03	79.59	71.93	46.37	48.05	1M text
ReVision-2M (ours)	✓	**74.94**	**36.40**	**76.35**	**45.23**	33.49	29.59	**26.80**	24.38	80.14	**72.18**	**48.26**	**49.75**	2M text

SFT dataset for this comparison, aligning exactly with the SFT data scale used in Unicorn (Yu et al., 2025). We compare four pre-training settings in Table 2: ① **Paired image-text:** 1M ground-truth image-text pairs (paired image-text baseline). ② **Unicorn:** 1M unpaired text samples, text-only method using mean shift. ③ **ReVision-1M**: 1M unpaired text samples. ④ **ReVision-2M:** 2M unpaired text samples. Comparing the text-only methods, ReVision-2M (49.75) dramatically outperforms Unicorn (43.94). Since Unicorn relies on simple mean-shifting, its synthesized features lack the correct geometric shape, leading to poor manifold penetration. ReVision generates pseudo-features that statistically mimic real visual distributions, indicating that the alignment strategy plays an important role in the quality of pseudo-visual representations. Additionally, we observe that Unicorn achieves notably high scores on Reasoning benchmarks. This discrepancy stems from the SFT data distribution: the InternVL-Chat-V1-2-SFT used in our controlled experiments focuses primarily on general visual conversation and basic perception, whereas Unicorn's original SFT dataset incorporates a massive volume of complex reasoning tasks. This suggests that ReVision's performance could be further elevated with more challenging instruction tuning data. Importantly, ReVision-2M (49.75) surpasses the paired image-text baseline (48.91) trained on 1M real paired samples, despite using only unpaired text during Stage-1 pretraining. This highlights a promising scaling way.

7 Related Works

The Modality Gap refers to the systematic offset between semantically corresponding image and text representations in a shared embedding space (Liang et al., 2022). C^3 (Zhang et al., 2024) further formalizes this phenomenon as a combination of a global displacement and isotropy residuals. Subsequent studies (Li et al., 2023a; Tewel et al., 2022; Su et al., 2022) have largely followed this framework to correct cross-modal discrepancies. Some works (Yu et al., 2025) reduce reliance on large-scale paired image-text data by mapping text representations into pseudo-visual supervision. However, these approaches mainly rely on simple mean shifting (Role et al., 2025), implicitly assuming a relatively simple residual structure, which limits their ability to match the more complex geometry of real visual representations (Liu et al., 2026). Accurately characterizing the geometric structure of the Modality Gap therefore remains a key challenge for high-quality modality alignment.

8 Conclusion

We analyze the Modality Gap in a fixed geometric frame, revealing stable biases, anisotropic residuals, and shared dominant geometry across modalities. Based on these findings, we propose ReAlign, a training-free statistical alignment method, and ReVision, a scalable paradigm for MLLM pretraining with unpaired text. Experiments show that ReVision consistently reduces cross-modal misalignment, providing an efficient path for scaling MLLMs.

AI USE STATEMENT

In this work, we used generative AI tools to assist with data processing, improve the clarity and presentation of the manuscript, and translate text. AI-assisted outputs were manually reviewed and revised by the authors to ensure correctness, consistency, and faithfulness to the intended meaning. The authors take full responsibility for the final content of the paper, including all text, claims, analyses, and artifacts produced with the aid of generative AI.

REFERENCES

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 7346–7370, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024c.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024d.

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

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, Jim Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *Advances in Neural Information Processing Systems*, 38:48009–48036, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024.

Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Chunyu Wang, Xiyang Dai, Dongdong Chen, et al. Llm2clip: Powerful language model unlocks richer visual representation. *arXiv preprint arXiv:2411.04997*, 2024.

Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*, 2023a.

Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Lingyu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *Advances in Neural Information Processing Systems*, 37:18535–18556, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023b.

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

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, 2026. URL https://arxiv.org/abs/2604.00279.

Yang Liu, Xiaomin Yu, Gongyu Zhang, Zhen Zhu, Christos Bergeles, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. Arcsin: Adaptive ranged cosine similarity injected noise for language-driven visual tasks. *arXiv preprint arXiv:2402.17298*, 2024.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*, 2022.

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

François Role, Sébastien Meyer, and Victor Amblard. Fill the gap: Quantifying and reducing the modality gap in image-text representation learning, 2025. URL https://arxiv.org/abs/2505.03703.

Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022.

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

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17918–17928, 2022.

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

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. In *European Conference on Computer Vision*, pp. 471–490. Springer, 2024.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Can Yaras, Siyi Chen, Peng Wang, and Qing Qu. Explaining and mitigating the modality gap in contrastive multimodal learning. *arXiv preprint arXiv:2412.07909*, 2024.

Lingjie Yi, Raphael Douady, and Chao Chen. Decipher the modality gap in multimodal contrastive learning: From convergent representations to pairwise alignment. *arXiv preprint arXiv:2510.03268*, 2025.

Xiaomin Yu, Pengxiang Ding, Wenjie Zhang, Siteng Huang, Songyang Gao, Chengwei Qin, Kejian Wu, Zhaoxin Fan, Ziyue Qiao, and Donglin Wang. Unicorn: Text-only data synthesis for vision language model training. *arXiv preprint arXiv:2503.22655*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Yuhui Zhang, Elaine Sui, and Serena Yeung-Levy. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. *arXiv preprint arXiv:2401.08567*, 2024.

A Additional Analysis of ReAlign

This section provides additional analysis of the statistical operations used in ReAlign. Let $e_y, e_x \in \mathbb{R}^d$ denote the source- and target-modality embeddings, with means $\mu_y = \mathbb{E}[e_y]$ and $\mu_x = \mathbb{E}[e_x]$, respectively. We denote their centered covariance matrices by $\Sigma_y = \mathrm{Cov}(e_y)$ and $\Sigma_x = \mathrm{Cov}(e_x)$.

A.1 Anisotropy-Preserving Calibration

ReAlign first removes the first-order displacement by translating the source representation to the target centroid:

$$\dot{e}_y = e_y - \mu_y + \mu_x.$$

By linearity of expectation, $\mathbb{E}[\dot{e}_y] = \mu_x$. Since translation does not affect centered covariance, we also have $\mathrm{Cov}(\dot{e}_y) = \Sigma_y$.

ReAlign then calibrates the global second-order scale. Let $T_y = \mathrm{tr}(\Sigma_y)$ and$T_x = \mathrm{tr}(\Sigma_x)$. The trace-matching scale is defined as:

$$s = \sqrt{\frac{T_x}{T_y + \varepsilon}}, \qquad \tilde{e}_y = \mu_x + s(e_y - \mu_y),$$

where $\varepsilon > 0$ is a small constant for numerical stability.

The centered covariance after scaling is:

$$\mathrm{Cov}(\tilde{e}_y) = s^2 \Sigma_y.$$

Therefore, Trace Alignment changes only the global variance scale while leaving the covariance eigenvectors unchanged. Moreover, the trace-normalized eigenspectrum is exactly preserved:

$$\frac{\lambda_j(\mathrm{Cov}(\tilde{e}_y))}{\mathrm{tr}(\mathrm{Cov}(\tilde{e}_y))} = \frac{\lambda_j(\Sigma_y)}{\mathrm{tr}(\Sigma_y)}.$$

Thus, the scaling operation preserves the relative anisotropic variance structure of the source modality while calibrating its total variance toward that of the target modality. This property applies to the linear scaling step itself; the subsequent sample-wise spherical normalization is nonlinear and does not generally preserve these statistics exactly.

A.2 Phantom Drift under Spherical Normalization

The translated and scaled representation does not generally satisfy the unit-norm constraint. ReAlign therefore projects it back onto the unit hypersphere:

$$e'_y = \frac{\tilde{e}_y}{\|\tilde{e}_y\|_2}.$$

Although $\mathbb{E}[\tilde{e}_y] = \mu_x$, spherical normalization is nonlinear, and expectation does not generally commute with normalization. Consequently, the centroid after projection, $\mu'_y = \mathbb{E}[e'_y]$, may differ from the target centroid μ_x. We define this normalization-induced displacement as the Phantom Drift:

$$\Delta_y^{\pi} = \mu'_y - \mu_x.$$

This explains why centroid alignment established before spherical projection may not be preserved after normalization and motivates the final correction step in ReAlign.

A.3 Centroid Correction under Norm Relaxation

ReAlign removes the Phantom Drift through a final translation:

$$e''_y = e'_y - \mu'_y + \mu_x.$$

The correction exactly restores the target centroid and leaves the centered covariance unchanged:

$$\mathbb{E}[e''_y] = \mu_x, \qquad \mathrm{Cov}(e''_y) = \mathrm{Cov}(e'_y).$$

Therefore, the final correction modifies only the first-order location of the projected distribution and does not alter its centered second-order geometry.

The final translation does not strictly preserve the unit norm. Let $h = \mu_x - \mu'_y$, such that $e''_y = e'_y + h$. Since $\|e'_y\|_2 = 1$, the reverse triangle inequality gives

$$\left| \|e''_y\|_2 - 1 \right| \leq \|h\|_2 = \|\mu_x - \mu'_y\|_2.$$

Hence, the norm deviation introduced by the final correction is bounded by the magnitude of the Phantom Drift itself. This provides a direct characterization of the trade-off between exact centroid restoration and strict unit-norm preservation.

B Distribution-Level Geometric Evaluation

We further evaluate ReAlign from a distribution-level geometric perspective. The goal is to examine whether ReAlign corrects the major cross-modal statistical mismatches while preserving the anisotropic structure already shared between image and text representations. We conduct the analysis on six vision-language encoders: LLM2CLIP, MetaCLIP 2, OpenCLIP-2B, OpenCLIP-80M, OpenCLIP-400M, and SigLIP 2. For each encoder, we reserve 90% of the representations for estimating ReAlign statistics and use the remaining 10% as a disjoint held-out evaluation set, resulting in approximately 900K calibration samples and 100K evaluation samples per encoder. Centroid and trace statistics are evaluated on the full held-out split. For covariance eigenspectra, dominant-subspace overlap, and angular analysis, we retain a 20K-sample held-out reservoir for computational efficiency. Angular distributions are estimated from 200K randomly sampled pairs, while the shared-PCA visualization uses 1,500 samples per distribution. No image-text correspondence is used to estimate the ReAlign transformation. For comparison, we additionally evaluate C3-style isotropic corruption with Gaussian noise followed by spherical renormalization, using the same held-out representations.

B.1 Stage-wise Geometric Analysis

We first examine how each stage of ReAlign affects the first- and second-order statistics of the source representation. Specifically, we evaluate five consecutive states: Original Text, Centroid Alignment, Trace Alignment, Spherical Projection, and Centroid Correction.

For first-order alignment, we measure the distance between the source and target centroids:

$$D_\mu = \|\mu_y - \mu_x\|_2 .$$

For the global second-order scale, we measure the relative covariance-trace mismatch:

$$D_{\mathrm{tr}} = \left| \frac{\mathrm{tr}(\Sigma_y)}{\mathrm{tr}(\Sigma_x)} - 1 \right| .$$

Fig. 8 shows a highly consistent first-order pattern across all six encoders. Before alignment, the centroid discrepancy ranges from approximately 0.39 to 1.03. Centroid Alignment reduces this discrepancy to approximately 2.7×10^{-3}–3.8×10^{-3} on held-out data. Trace Alignment leaves the centroid essentially unchanged, consistent with its construction around the target centroid.

Spherical projection subsequently increases the centroid discrepancy for every encoder. The effect is particularly pronounced for SigLIP 2, where the discrepancy increases to approximately 6.2×10^{-2}. This provides direct empirical evidence for the normalization-induced Phantom Drift analyzed in Sec. A. The final Centroid Correction reduces the held-out discrepancy again to approximately 3×10^{-3}.

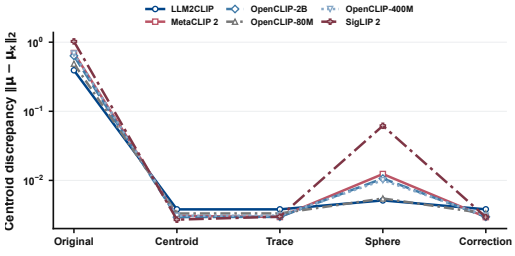


Figure 8: Stage-wise centroid discrepancy across six encoders. Centroid Alignment removes most of the initial cross-modal displacement, while Trace Alignment preserves the corrected centroid. Spherical projection introduces a new displacement, demonstrating Phantom Drift, and the final Centroid Correction restores first-order alignment on held-out data.

Because all alignment statistics are estimated from the calibration split and evaluated on disjoint held-out samples, the final centroid discrepancy does not vanish exactly. The remaining error reflects finite-sample generalization of the estimated modality-level statistics.

The trace results in Fig. 9 show a complementary pattern. Centroid translation leaves the trace mismatch unchanged, as expected from translation invariance of centered covariance. Trace Alignment reduces the relative mismatch to between approximately 1.4×10^{-5} and 4.9×10^{-4} across all six encoders.

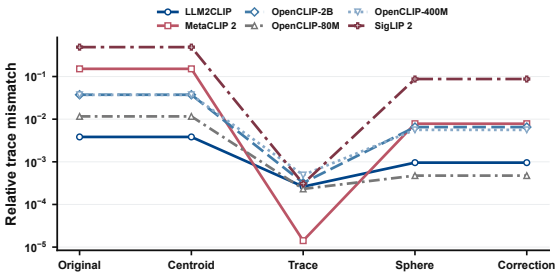


Figure 9: **Stage-wise relative covariance-trace mismatch.** Centroid translation leaves the centered second-order scale unchanged. Trace Alignment sharply reduces the global variance mismatch. Spherical projection modifies the trace because of its nonlinear sample-wise operation, while the final centroid translation leaves the post-projection covariance trace unchanged.

Spherical projection subsequently modifies the global second-order scale. The resulting change is mild for most encoders but more pronounced for SigLIP 2, where the held-out trace mismatch increases to approximately 8.7×10^{-2}. The final Centroid Correction leaves the trace unchanged from the spherical-projection stage because translation does not affect centered covariance.

These results verify the intended role of each component of ReAlign: centroid translation corrects the dominant first-order displacement, Trace Alignment accurately calibrates the global variance scale before spherical projection, spherical normalization introduces a new centroid drift, and the final translation restores first-order alignment without modifying the post-projection centered covariance.

B.2 Spectral Structure Preservation

We next examine whether ReAlign preserves the anisotropic variance structure identified in Sec. 3.3. For a covariance matrix Σ with eigenvalues $\lambda_j(\Sigma)$, we define its trace-normalized eigenspectrum as

$$p_j = \frac{\lambda_j(\Sigma)}{\mathrm{tr}(\Sigma)}.$$

Fig. 10 compares the normalized eigenspectra of Image Target, Original Text, C3, and ReAlign. Across all six encoders, the image and original text representations exhibit strongly non-uniform spectral decay, indicating that anisotropic second-order structure is already shared across modalities.

For five of the six encoders, ReAlign almost completely preserves the source spectral profile. The ℓ_1 discrepancy between the normalized spectra of ReAlign and Original Text remains below 0.009 for LLM2CLIP, MetaCLIP 2, OpenCLIP-2B, OpenCLIP-80M, and OpenCLIP-400M. SigLIP 2 exhibits a larger change of approximately 0.086, although its representation remains strongly anisotropic.

In contrast, the C3 pipeline produces substantially larger spectral changes. Its normalized-spectrum discrepancy from the original text representation ranges from approximately 0.57 to 0.94 across the six encoders. The resulting spectra are considerably flatter and move toward a more uniform allocation of variance across directions.

These results support the use of global scalar calibration in ReAlign. The method preserves the relative variance structure already present in the source representation instead of replacing it with an approximately isotropic perturbation.

B.3 Dominant-Subspace Preservation

Similar eigenvalue profiles do not necessarily imply similar dominant directions. We therefore additionally evaluate the overlap between the leading covariance eigenspaces.

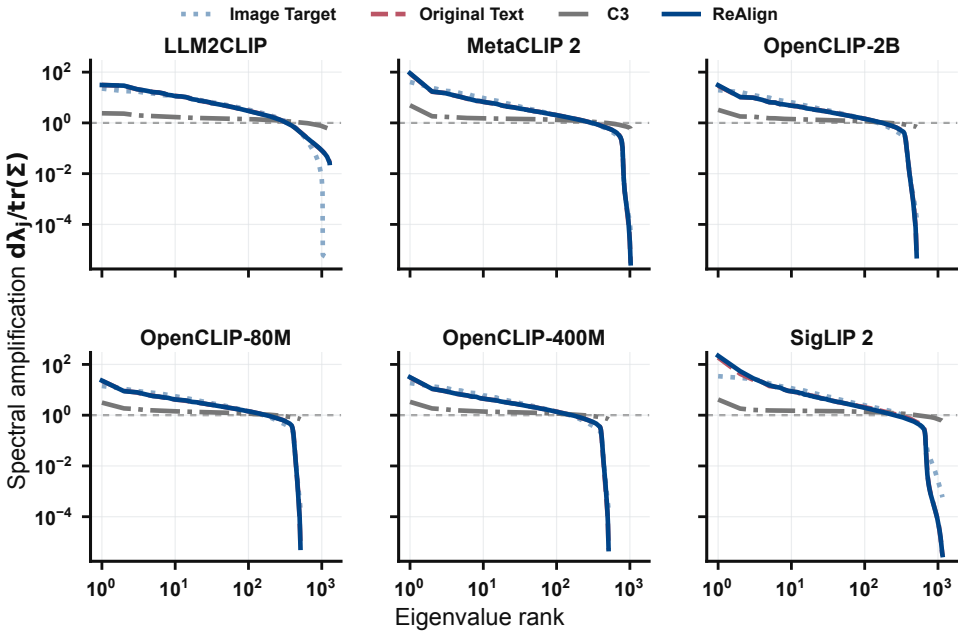


Figure 10: **Trace-normalized covariance eigenspectra across six encoders.** Image and text representations exhibit similar anisotropic spectral decay. ReAlign closely preserves the source spectral structure for five encoders and retains a strongly anisotropic profile for SigLIP 2. The C3 pipeline produces substantially flatter spectra. The horizontal dashed line denotes the isotropic reference level.

Let $U_x^{(q)}$ and $U_y^{(q)}$ contain the top-q covariance eigenvectors of the image and evaluated representations, respectively. We measure their dominant-subspace overlap as

$$S_q = \frac{1}{q} \left\| U_x^{(q)\top} U_y^{(q)} \right\|_F^2 .$$

We report $q \in \{32, 64, 128\}$. A larger value indicates greater overlap between the corresponding dominant subspaces. As shown in Fig. 11, ReAlign closely follows the overlap of Original Text for five of the six encoders across all evaluated subspace dimensions. For example, at $q = 128$, ReAlign achieves overlaps of approximately 0.44, 0.48, 0.69, 0.63, and 0.62 for LLM2CLIP, MetaCLIP 2, OpenCLIP-2B, OpenCLIP-80M, and OpenCLIP-400M, respectively, closely matching their original text representations. SigLIP 2 exhibits a larger reduction after alignment. Its overlap decreases from approximately 0.38 to 0.33 at $q = 128$, consistent with the stronger normalization-induced changes observed in the previous analyses. Nevertheless, its overlap remains substantially higher than that of C3.

The C3 pipeline consistently produces substantially lower dominant-subspace overlap. Since the evaluated C3 pipeline combines isotropic corruption with subsequent spherical renormalization, the reported reduction reflects the complete transformation rather than Gaussian perturbation alone.

These results clarify the role of ReAlign. Its objective is to retain the dominant geometry already shared across modalities while correcting the statistical mismatches required for alignment. The method therefore does not require explicit reconstruction or rotation of the full target covariance structure.

B.4 Angular Geometry Preservation

Covariance statistics characterize only part of the representation geometry. We therefore further examine sample-level relational structure through pairwise angular distributions.

For each representation, we randomly sample pairs of held-out embeddings and compute their cosine similarities. Let P_x denote the empirical cosine-similarity distribution of image representations

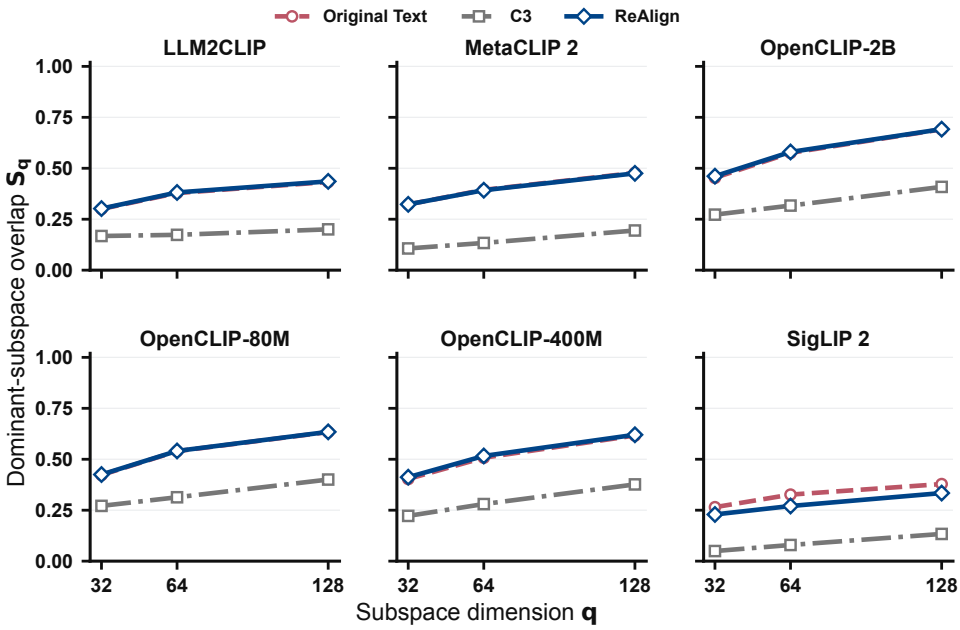


Figure 11: **Dominant-subspace overlap with image representations.** ReAlign closely preserves the overlap already present in Original Text for five of the six encoders and remains substantially above C3 for all evaluated encoders. SigLIP 2 exhibits a larger post-alignment reduction, consistent with its stronger normalization-induced drift.

and P_y denote that of an evaluated representation. We quantify their discrepancy using the Jensen–Shannon divergence:

$$D_{\mathrm{JS}} = \mathrm{JS}\left(P_x \,\|\, P_y\right).$$

Fig. 12 shows that ReAlign yields substantially lower angular divergence than C3 for all six encoders. The ReAlign divergence ranges from approximately 0.002 to 0.178, whereas C3 ranges from approximately 0.249 to 0.689.

ReAlign also improves over Original Text for five of the six encoders. The only exception is LLM2CLIP, for which the original image and text angular distributions are already extremely similar: the JS divergence increases slightly from approximately 0.0011 to 0.0065 after ReAlign while remaining far below the C3 value of approximately 0.279.

The results show that the C3 pipeline substantially shifts the pairwise angular distribution, whereas ReAlign produces an angular distribution much closer to that of the image representations while correcting the dominant cross-modal displacement.

Together with the spectral and subspace analyses, these results show that ReAlign preserves source-side second-order structure while achieving substantially better agreement with the target angular distribution.

B.5 Shared-PCA Visualization

We additionally visualize the distributions in a common low-dimensional coordinate system. For each encoder, Image, Original Text, C3, and ReAlign representations are jointly used to fit a single PCA basis:

$$[X,\, Y,\, Y_{\mathrm{C3}},\, Y_{\mathrm{ReAlign}}] \longrightarrow U_{\mathrm{PCA}}.$$

All four distributions are then projected using the same PCA center and basis. Using one shared projection is important because independently fitted PCA coordinate systems cannot be directly compared.

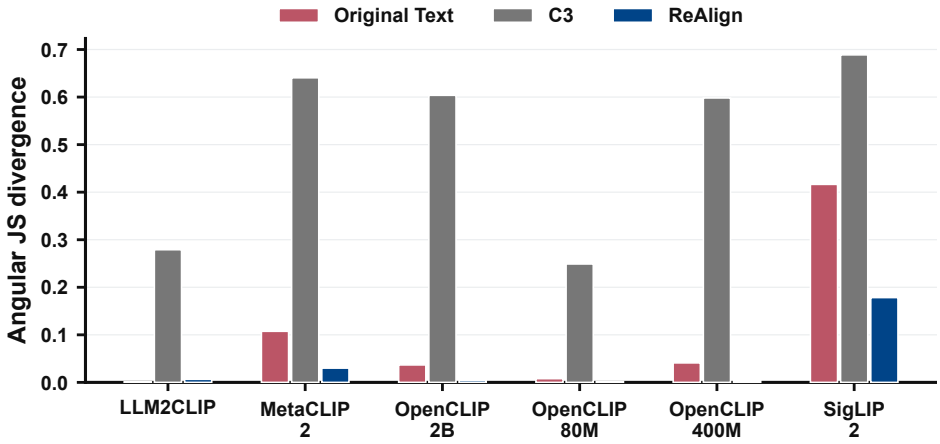


Figure 12: **Pairwise angular-distribution discrepancy.** Jensen–Shannon divergence is measured between the cosine-similarity distributions of image representations and each evaluated representation. ReAlign exhibits substantially smaller divergence than C3 for all six encoders and improves over Original Text for five encoders.

Fig. 13 provides a qualitative view consistent with the quantitative results above. The original text representations exhibit substantial displacement from the image distributions. C3 moves the source representation toward the image region but produces a comparatively compact and more isotropic-looking cloud. ReAlign substantially reduces the visible modality displacement while retaining pronounced anisotropic variation.

The aligned representation is not expected to reproduce the complete image covariance orientation. ReAlign uses global scalar calibration rather than full covariance rotation, and its design explicitly retains the useful source geometry already shared with the target modality. The PCA visualization is therefore used as qualitative evidence and complements the spectral, subspace, and angular metrics above.

B.6 Norm Deviation after Centroid Correction

The final Centroid Correction is a Euclidean translation and therefore does not strictly preserve the unit-norm constraint established by spherical projection. We quantify its effect using the mean absolute norm deviation:

$$D_{\text{norm}} = \mathbb{E}\left[\left|\|e''_y\|_2 - 1\right|\right].$$

Fig. 14 shows that the resulting deviation is small for most evaluated encoders. The mean deviation is approximately 6.2×10^{-4} for LLM2CLIP, 6.5×10^{-4} for OpenCLIP-80M, and between approximately 2.3×10^{-3} and 3.4×10^{-3} for MetaCLIP 2, OpenCLIP-2B, and OpenCLIP-400M. SigLIP 2 exhibits a larger mean deviation of approximately 3.0×10^{-2}, consistent with its stronger Phantom Drift.

More importantly, the empirical deviations satisfy the bound derived in Sec. A. For all six encoders, the observed maximum deviation remains below the norm of the final centroid-correction vector. For example, SigLIP 2 has a calibration Phantom Drift magnitude of approximately 0.0615 and a maximum observed norm deviation of approximately 0.0547.

Thus, the final translation produces a bounded departure from the unit hypersphere while restoring first-order alignment and leaving the centered covariance unchanged.

C Robustness, Efficiency, and Domain Sensitivity

We further evaluate the practical robustness of ReAlign from three complementary perspectives: ① the amount of calibration data required to obtain stable alignment statistics, ② numerical accuracy and computational scalability, and ③ sensitivity to distribution shifts between the calibration and evaluation domains.

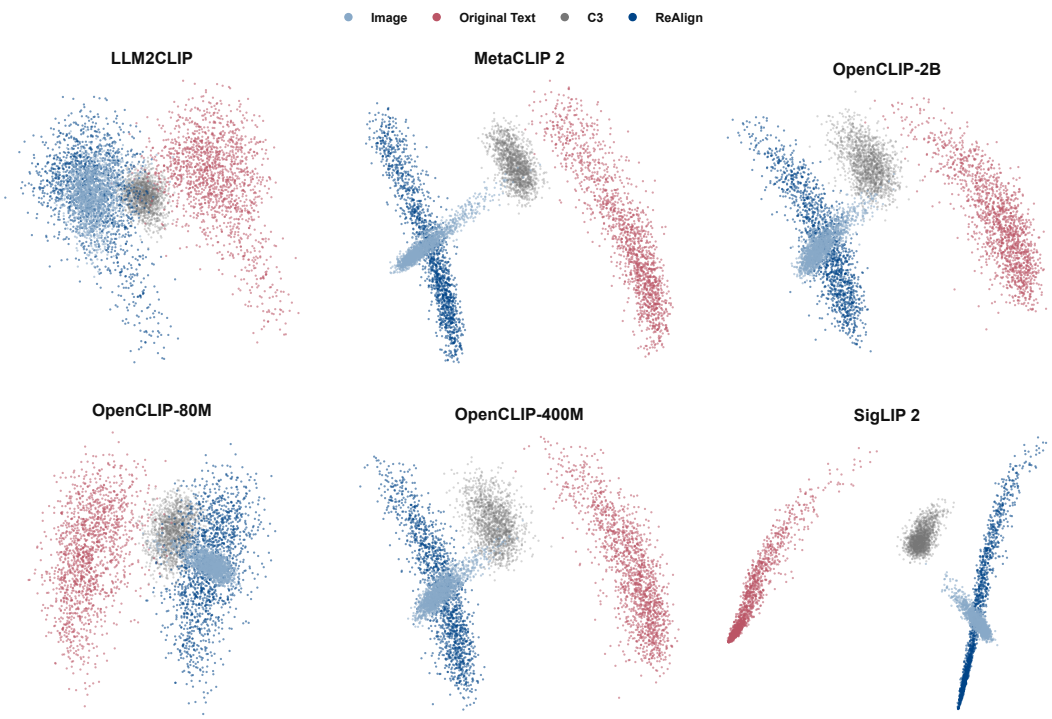


Figure 13: **Shared-PCA visualization across six encoders.** For each encoder, Image, Original Text, C3, and ReAlign are jointly projected using a single PCA basis. ReAlign substantially reduces the visible cross-modal displacement while retaining anisotropic structure, whereas C3 forms a more compact isotropic-like distribution. PCA is used only for qualitative visualization.

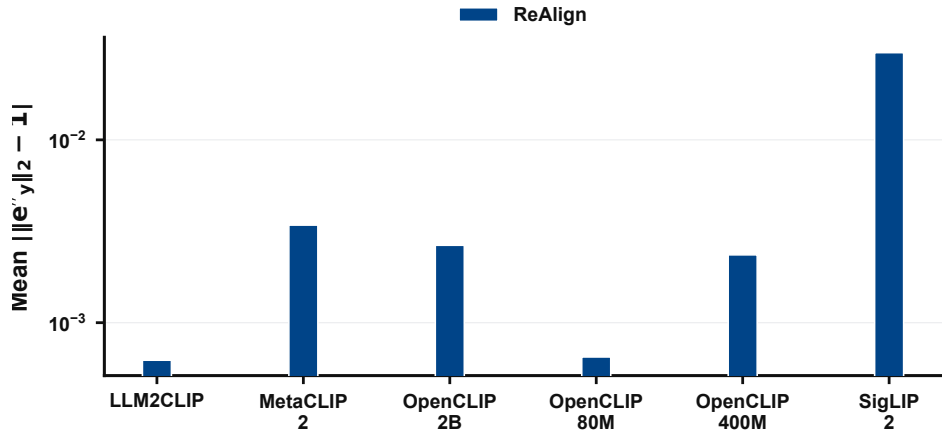


Figure 14: **Mean norm deviation after the final Centroid Correction.** The final translation does not strictly preserve unit norm, but the resulting deviation remains small for most encoders. SigLIP 2 exhibits a larger deviation, consistent with its stronger Phantom Drift. The maximum observed deviation remains within the theoretical correction bound for all six encoders.

C.1 Calibration Data Efficiency

ReAlign relies only on a small set of modality-level statistics estimated from unpaired data. Specifically, the transformation requires modality centroids, global covariance traces, and the post-normalization centroid used by the final correction. It therefore avoids iterative optimization and does not require estimating or storing a full covariance matrix.

To evaluate the sample efficiency of these statistics, we randomly sample N unpaired image and text representations from the calibration corpus, where

$$N \in \left\{10^2,\, 5\times10^2,\, 10^3,\, 5\times10^3,\, 10^4,\, 5\times10^4,\, 10^5,\, 5\times10^5\right\}.$$

For each N, all ReAlign calibration statistics are estimated exclusively from the sampled subsets. The resulting transformation is then evaluated on a fixed held-out set. We report the mean and standard deviation over $K = 5$ independent trials.

As shown in Fig. 15, the alignment result improves rapidly as the calibration set increases and becomes stable once the sample size reaches approximately 10^4.

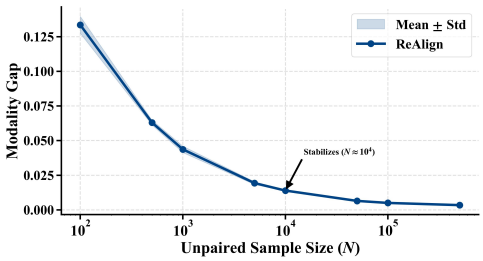


Figure 15: Impact of unpaired sample size on the modality gap for statistical estimation.

Observation. ① **Rapid statistical convergence.** The largest variation occurs in the low-sample regime, while both the mean performance and the variance across independent trials stabilize after a moderate number of calibration samples. This behavior is consistent with ReAlign relying on low-order population statistics rather than parameters learned through iterative optimization.

② **Practical calibration efficiency.** The results indicate that accurate estimation of the required modality-level statistics does not require the full pretraining corpus. A relatively small unpaired calibration subset is sufficient to obtain stable alignment, making ReAlign practical when only limited data from a target distribution are available.

C.2 Numerical Stability and Computational Complexity

We next examine whether streaming estimation of ReAlign statistics remains numerically reliable and computationally scalable for large embedding collections.

When a large number of high-dimensional vectors are accumulated, finite-precision arithmetic can introduce numerical error in the estimated statistics. We therefore compare `Float32` and `Float64` accumulation using the high-precision estimate as the numerical reference.

As shown in Fig. 16(a), the error accumulated with `Float32` increases with the number of processed samples and reaches approximately 9.56×10^{-8} at $N = 500{,}000$. Using `Float64` accumulation reduces this numerical residual by several orders of magnitude and keeps it close to double-precision numerical accuracy.

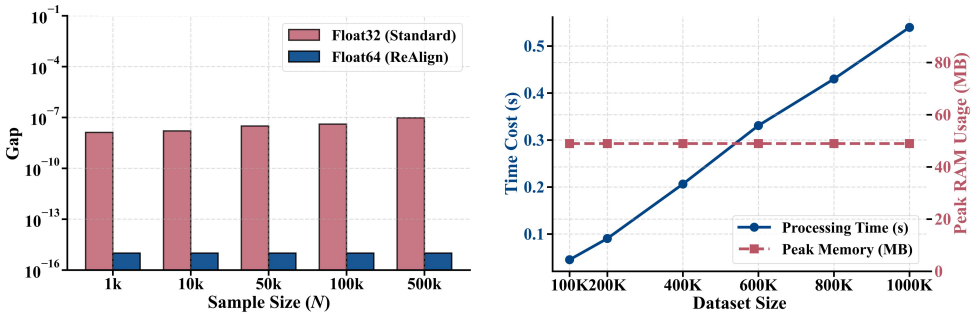
Although the absolute `Float32` error remains small, these results show that double-precision accumulation provides a simple way to prevent numerical error from accumulating when estimating population statistics from very large datasets. We therefore use `Float64` accumulators for statistical estimation in our implementation.

We additionally analyze runtime and memory consumption by increasing the number of processed embeddings from 100K to 1M. For embedding dimension d, ReAlign requires only streaming accumulation of first-order statistics and scalar second-order quantities. Its computational complexity is therefore $O(Nd)$ in time and $O(d)$ in auxiliary memory.

For a fixed embedding dimension, this corresponds to linear scaling with dataset size and constant auxiliary memory with respect to N.

As shown in Fig. 16(b), the measured processing time grows approximately linearly from 0.05 s for 100K samples to 0.57 s for 1M samples. Meanwhile, peak memory usage remains approximately 48.95 MB over the evaluated range.

These results are consistent with the theoretical streaming complexity of ReAlign and show that its calibration cost scales favorably to large embedding collections without requiring storage of the full dataset in memory.



(a) Numerical Stability (b) Time & RAM

Figure 16: (a) Comparison of numerical residuals using Float32 and Float64 accumulators. (b) Processing time and peak memory usage across dataset sizes. For embedding dimension d, ReAlign requires $O(Nd)$ time and $O(d)$ auxiliary memory, corresponding to linear-time and constant-memory scaling with respect to N for fixed d.

C.3 Sensitivity to Domain Mismatch

The statistics used by ReAlign characterize the relationship between two modalities under a particular data distribution. A natural question is therefore whether statistics estimated from one domain can be directly transferred to another domain.

We evaluate this using two substantially different data distributions: ① **General**: Bunny-Pretrain, representing a general-domain image-text distribution; and ② **Medical**: PubMedVision-Pretrain Chen et al. (2024a), representing a specialized biomedical distributi

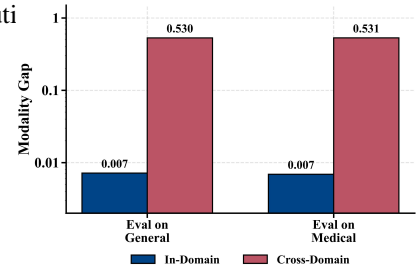


Figure 17: Comparison of modality gap performance under in-domain and cross-domain statistical alignment for General and Medical domains.

Protocol. Let $\mathcal{D}_{\text{stat}}$ denote the domain used to estimate the ReAlign statistics and $\mathcal{D}_{\text{eval}}$ denote the domain on which the resulting transformation is evaluated. We consider four combinations: General $\rightarrow$ General, General $\rightarrow$ Medical, Medical $\rightarrow$ Medical, and Medical $\rightarrow$ General.

For every setting, the transformation is estimated only from $\mathcal{D}_{\text{stat}}$ and evaluated on held-out representations from $\mathcal{D}_{\text{eval}}$.

As shown in Fig. 17, statistics estimated and evaluated within the same domain consistently produce smaller modality gaps than statistics transferred across the two domains. Using General-domain statistics on Medical representations results in a larger residual gap than Medical $\rightarrow$ Medical calibration. The reverse transfer from Medical to General similarly degrades alignment relative to General $\rightarrow$ General.

Observation. These results indicate that the modality-level statistics corrected by ReAlign are sensitive to the underlying data distribution. In particular, changes in semantic domain can alter modality centroids, global variance scales, and the normalization-induced correction required by the alignment transformation.

Consequently, when the deployment distribution differs substantially from the calibration distribution, re-estimating ReAlign statistics on unpaired samples from the target domain can provide more accurate alignment. Combined with the sample-efficiency results above, this recalibration requires only a moderate amount of target-domain data and does not involve additional optimization.

D Failure Analysis: Blockwise Covariance Alignment

A natural alternative to ReAlign is to perform more fine-grained second-order matching by explicitly transforming the source covariance toward the target covariance. We investigate this possibility through a training-free baseline, termed **Blockwise Covariance Alignment (BC Align)**. The purpose

of this experiment is to test whether stronger covariance matching provides additional benefits over the scalar trace calibration adopted by ReAlign.

BC Align uses only statistics estimated from large unpaired image and text collections. It first performs centroid translation, then applies whitening–coloring transforms to covariance blocks, followed by spherical projection and centroid correction. Compared with ReAlign, BC Align therefore uses substantially more detailed second-order statistics and imposes a stronger geometric transformation on the source representation.

D.1 Blockwise Covariance Alignment

Let $e_a, e_b \in \mathbb{R}^d$ denote unit-normalized embeddings from source modality a and target modality b, with $\mu_a = \mathbb{E}[e_a]$, and $\mu_b = \mathbb{E}[e_b]$. We consider a fixed orthogonal decomposition:

$$\mathbb{R}^d = U \oplus V,$$

with corresponding projectors P_U and P_V. BC Align performs three stages.

Centroid translation. The source representation is first translated toward the target centroid:

$$e_a^{(1)} = e_a - \mu_a + \mu_b.$$

This Euclidean translation matches the population centroid before normalization. When spherical projection is subsequently applied, the centroid may shift again because normalization is nonlinear.

Blockwise covariance matching. Using the translated source embeddings, we estimate the blockwise covariance matrices:

$$\Sigma_{U,a} = \mathrm{Cov}(P_U e_a^{(1)}), \qquad \Sigma_{V,a} = \mathrm{Cov}(P_V e_a^{(1)}),$$

and analogously obtain $\Sigma_{U,b}$ and $\Sigma_{V,b}$ from the target modality. We then define regularized whitening–coloring transformations:

$$T_{U,a} = (\Sigma_{U,b} + \epsilon I)^{1/2}(\Sigma_{U,a} + \epsilon I)^{-1/2}, \quad T_{V,a} = (\Sigma_{V,b} + \epsilon I)^{1/2}(\Sigma_{V,a} + \epsilon I)^{-1/2},$$

and construct:

$$e_a^{(2)} = T_{U,a} P_U e_a^{(1)} + T_{V,a} P_V e_a^{(1)}.$$

Before spherical projection, these transformations match the corresponding covariance blocks under the regularized whitening–coloring objective. The representation is then projected onto the unit sphere:

$$\hat{e}_a = \frac{e_a^{(2)}}{\|e_a^{(2)}\|_2}.$$

Because spherical projection is nonlinear, the covariance after normalization is not guaranteed to remain equal to the target covariance.

Centroid correction. Let $\hat{\mu}_a = \mathbb{E}[\hat{e}_a]$ denote the post-normalization centroid. As in ReAlign, spherical projection may introduce a new centroid displacement. We therefore apply a final translation:

$$e_a^{\mathrm{BC}} = \hat{e}_a - \hat{\mu}_a + \mu_b.$$

No additional normalization is applied after this final correction, since doing so would generally introduce another centroid drift.

D.2 Why Does Fine-Grained Covariance Alignment Underperform?

Although BC Align uses a substantially more expressive second-order transformation than ReAlign, it consistently yields lower downstream performance in our experiments. Table 3 compares the two approaches across general, reasoning, and hallucination-oriented benchmarks.

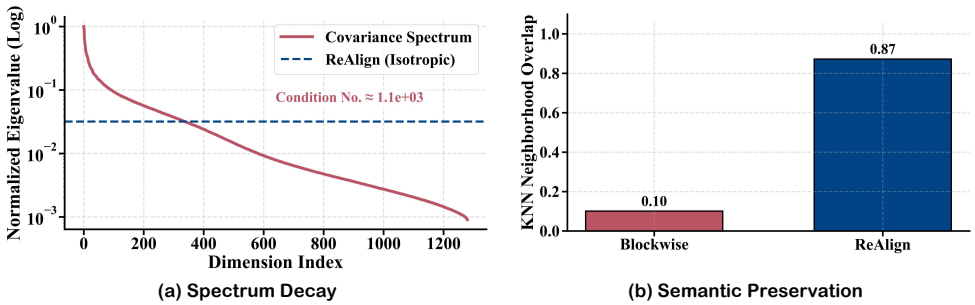


Figure 18: **Failure mechanism analysis of Blockwise Covariance Alignment.** (a) **Spectrum Decay:** The large condition number ($\approx 1.1 \times 10^3$) of the text embedding covariance makes inverse-square-root estimation sensitive to low-variance tail directions, where small estimation errors can be strongly amplified. ReAlign avoids this matrix inversion and instead uses stable scalar trace calibration. (b) **Local Structure Preservation:** KNN neighborhood overlap shows that Blockwise Covariance Alignment substantially alters the source local geometry, retaining only about 10% of the original neighborhoods, whereas ReAlign preserves approximately 87%.

Table 3: **Performance comparison between BC Align and ReAlign.** ReAlign consistently outperforms the more fine-grained blockwise covariance matching baseline across the evaluated benchmarks, suggesting that stronger second-order matching does not necessarily translate into better downstream representations.

Method	General				Reasoning				Hallucination			Avg.
	MME	MMStar	SQA	RealWorldQA	MMMU	MMMU-P	VisuLogic	LogicVista	CRPE	POPE	HallBench	
BC Align	76.12	34.33	76.42	45.36	30.69	27.95	24.40	22.60	80.93	71.08	46.27	48.74
ReAlign	**79.65**	**36.13**	**76.71**	**47.97**	**31.51**	**28.39**	**27.70**	**22.82**	**81.78**	**72.53**	**46.58**	**50.16**

Sensitivity of full covariance estimation. ReAlign requires only a scalar second-order statistic,

$$T = \mathrm{tr}(\Sigma),$$

which can be estimated robustly from large unpaired samples. BC Align instead requires estimating matrix-valued covariance statistics and applying inverse square-root transformations. As shown in Fig. 18(a), the empirical covariance spectrum is strongly anisotropic and contains small tail eigenvalues, resulting in a condition number of approximately $\kappa \approx 1.10 \times 10^3$. For a whitening transform, a direction with eigenvalue λ_i is scaled proportionally to $\lambda_i^{-1/2}$. Consequently, estimation errors associated with low-variance directions can be substantially amplified by covariance inversion. Although regularization alleviates singularity, the resulting transformation remains considerably more sensitive to tail-spectrum estimation than the scalar trace calibration used by ReAlign.

Distortion from aggressive covariance reshaping. The distribution-level analysis in Sec. B shows that image and text representations already share substantial spectral anisotropy and dominant-subspace structure. This suggests that fully reshaping the source covariance is unnecessary and may modify useful geometry that is already shared across modalities. ReAlign adopts a conservative transformation:

$$\tilde{e}_y = \mu_x + s(e_y - \mu_y),$$

which preserves the covariance eigendirections and trace-normalized spectrum before spherical projection. BC Align, in contrast, applies direction-dependent transformations in order to match the target covariance blocks. We quantify the resulting local distortion using k-nearest-neighbor overlap with $k = 10$. As shown in Fig. 18(b), BC Align preserves only 10.1% of the original local neighborhoods, whereas ReAlign retains 87.3%. Thus, the stronger covariance transformation substantially alters the local neighborhood structure of the source representation. These observations are consistent with the downstream results in Table 3. More expressive covariance matching provides a closer second-order fitting objective, but it also introduces greater estimation sensitivity and stronger distortion of the source geometry. ReAlign instead uses robust first-order statistics and a scalar

second-order calibration, preserving the shared geometric structure while correcting the dominant cross-modal statistical mismatch.

E EXPERIMENTS SETTING

E.1 MLLM TRAINING SETTING

We evaluate three vision-language backbone configurations: LLM2CLIP-OpenAI-L-14-336 with Qwen3-8B, SigLIP2-So400m-patch14-384 with Qwen3-8B, and LLM2CLIP-OpenAI-L-14-336 with Llama-3-8B-Instruct. The input modalities are connected to the LLM through a two-layer MLP projector with GELU activation. Our core design utilizes aligned text representations as pseudo-visual tokens, which are mapped into the LLM's feature space through the same projector used for visual representations. **Training Pipeline.** The training process consists of two stages: ① **Stage 1: Modality Substitution Pretraining.** We train the projector for 1 epoch on the filtered Bunny-1M dataset with a learning rate of 5×10^{-4}, while keeping the LLM parameters frozen. ② **Stage 2: Visual Instruction Tuning.** We perform full-parameter fine-tuning on the InternVL-Chat-V1.2 dataset for 1 epoch. The projector is initialized with weights from Stage 1, and the learning rate is reduced to 1×10^{-5}. The experiments were conducted on 8 NVIDIA H200 GPUs. With a total data scale of approximately 2.2M samples, the complete training pipeline was finished in 12 hours.

E.2 EVAL SETTING

We evaluate ReVision across three primary dimensions using a comprehensive suite of benchmarks: General Perception (MME test Fu et al. (2025), MMStar Chen et al. (2024c), ScienceQA-image dev&test Lu et al. (2022), and RealWorldQA), Complex Reasoning (MMMU validation single-image Yue et al. (2024), MMMU-Pro single-image Yue et al. (2025), VisuLogic train Xu et al. (2025), and LogicVista Xiao et al. (2024)), and Hallucination Assessment (CRPE Wang et al. (2024), POPE Li et al. (2023b), and HallusionBench Guan et al. (2024)). For all benchmarks, we report accuracy (acc) as the unified metric to ensure a consistent and rigorous comparison.

E.3 PRETRAINING DATA EFFICIENCY

We further compare different pretraining strategies from the perspective of Stage-1 data requirements. Rather than introducing a monetary cost estimate that depends on a particular API provider or pricing scheme, we directly report the type and scale of pretraining supervision used by each method.

We consider four controlled settings: ① **Paired Image–Text**, which uses 1M paired image-text samples for Stage-1 pretraining; ② **Unicorn**, which uses 1M unpaired text samples; ③ **ReVision-1M**, which uses 1M unpaired text samples; and ④ **ReVision-2M**, which scales the unpaired text corpus to 2M samples. To ensure a controlled comparison, all methods use the same 417K Stage-2 visual instruction tuning subset. Therefore, the primary difference lies in the type and scale of supervision used during Stage 1. As reported in Table 2, the paired image–text baseline achieves an average score of 48.91 using 1M paired samples. With the same Stage-1 sample scale, ReVision-1M achieves 48.05 using only unpaired text, approaching the paired-data baseline without requiring image-text correspondence. When the text-only pretraining corpus is increased to 2M samples, ReVision-2M further improves the average score to 49.75, surpassing the 1M paired image–text baseline.

In contrast, Unicorn achieves an average score of 43.94 with 1M unpaired text samples. This gap indicates that the effectiveness of text-only pretraining depends not only on the amount of text data, but also on how textual representations are mapped into the visual representation space. Overall, these results show that ReVision benefits from scaling unpaired text-only pretraining and that, under the controlled setting considered here, 2M unpaired text samples can outperform pretraining with 1M paired image-text samples. This suggests that effective cross-modal representation alignment can make large-scale unpaired text a viable source of multimodal pretraining supervision.

Table 4: Ablation of ReAlign components. Step 1 denotes Anchor Alignment, Step 2 denotes Trace Alignment, and the full ReAlign additionally includes Centroid Alignment.

Method	General				Reasoning				Hallucination			Avg. ↑
	MME	MMStar	SQA	RealWorldQA	MMMU	MMMU-P	VisuLogic	LogicVista	CRPE	POPE	HallBench	
w/. Step 1	74.86	33.18	73.93	42.64	28.48	25.61	22.94	18.86	78.93	70.82	43.24	46.68
w/. Step 1 & 2	78.22	35.71	76.21	45.16	30.05	26.09	26.05	22.16	80.37	71.18	44.14	48.67
ReAlign / ReVision	**79.65**	**36.13**	**76.71**	**47.97**	**31.51**	**28.39**	**27.70**	**22.82**	**81.78**	**72.53**	**46.58**	**50.16**

Table 5: Ablation of Modality Substitution Pretraining. Removing Stage 1 consistently degrades performance across all benchmarks.

Method	General				Reasoning				Hallucination			Avg. ↑
	MME	MMStar	SQA	RealWorldQA	MMMU	MMMU-P	VisuLogic	LogicVista	CRPE	POPE	HallBench	
w/o Stage 1	66.80	29.70	71.90	39.10	25.60	23.40	20.10	16.90	75.80	68.90	41.20	43.58
ReVision	**79.65**	**36.13**	**76.71**	**47.97**	**31.51**	**28.39**	**27.70**	**22.82**	**81.78**	**72.53**	**46.58**	**50.16**

F More MLLM Experiments

In this section, we provide additional MLLM experiments to further analyze the effectiveness and scalability of ReVision. We first ablate the three components of ReAlign, then study the role of Modality Substitution Pretraining, and finally evaluate scaling behavior with respect to Stage-2 visual instruction data and LLM backbone size.

F.1 Ablation of ReAlign Components

To verify the contribution of each step in ReAlign, we conduct a component-wise ablation by progressively adding Anchor Alignment, Trace Alignment, and Centroid Alignment. As shown in Table 4, Anchor Alignment alone already provides a strong baseline by correcting the first-order mean mismatch. Adding Trace Alignment further improves the average score by matching the global residual energy scale while preserving the learned spectral structure. Finally, Centroid Alignment brings additional gains by correcting the centroid drift induced by spherical normalization. These results support the design of ReAlign as a three-stage calibration procedure.

F.2 Effect of Modality Substitution Pretraining

We further ablate the first stage of ReVision to examine whether Modality Substitution Pretraining is necessary. In the w/o Stage 1 setting, the model skips pseudo-visual pretraining and directly proceeds to visual instruction tuning. As shown in Table 5, removing Stage 1 leads to a substantial performance drop across all benchmarks, decreasing the average score from 50.16 to 43.58. This confirms that Stage 1 is not merely a generic projector warm-up; instead, it provides a visually compatible semantic interface before real-image instruction tuning.

F.3 Scaling Visual Instruction Tuning Data

We study how the amount of Stage-2 visual instruction tuning data affects ReVision. We keep Stage 1 fixed and vary only the amount of Stage-2 SFT data. As shown in Table 6, increasing the SFT data from 30% to 60% and then to 100% consistently improves performance. This trend indicates that Stage 1 establishes a useful cross-modal semantic interface, while Stage 2 progressively injects fine-grained real-image supervision and instruction-following ability.

F.4 Scaling the LLM Backbone

We evaluate whether ReVision benefits from scaling the LLM backbone. We replace only the LLM backbone from Llama-3-8B-Instruct to Llama-3-70B-Instruct, while keeping the visual/text encoder, ReAlign operator, Stage-1 corpus, Stage-2 SFT dataset, and evaluation suite unchanged. As shown

Table 6: Scaling Stage-2 visual instruction tuning data. Stage 1 is fixed, and only the amount of Stage-2 SFT data is varied.

Stage-2 Data	General				Reasoning				Hallucination			Avg. ↑
	MME	MMStar	SQA	RealWorldQA	MMMU	MMMU-P	VisuLogic	LogicVista	CRPE	POPE	HallBench	
30%	68.53	30.60	70.90	41.88	28.67	25.33	23.60	17.17	74.65	66.28	40.35	44.36
60%	75.99	34.67	75.06	45.36	30.12	27.47	25.10	20.06	78.57	70.53	44.39	47.94
100%	**79.65**	**36.13**	**76.71**	**47.97**	**31.51**	**28.39**	**27.70**	**22.82**	**81.78**	**72.53**	**46.58**	**50.16**

Table 7: Scaling the LLM backbone under the same ReVision pipeline. Only the LLM backbone is changed, while all other components and data settings remain fixed.

Backbone	General				Reasoning				Hallucination			Avg. ↑
	MME	MMStar	SQA	RealWorldQA	MMMU	MMMU-P	VisuLogic	LogicVista	CRPE	POPE	HallBench	
Llama-3-8B-Instruct	79.65	36.13	76.71	47.97	31.51	28.39	27.70	22.82	81.78	72.53	46.58	50.16
Llama-3-70B-Instruct	**80.27**	**37.84**	**78.92**	**49.16**	**34.99**	**31.27**	**31.54**	**26.11**	**82.52**	**73.89**	**49.15**	**52.33**

in Table 7, scaling the backbone improves the average score from 50.16 to 52.33. The gains are especially visible on reasoning-intensive benchmarks, suggesting that once the geometric pseudo-visual interface is established, a larger LLM can better exploit the injected semantic supervision.

The Modality Gap refers to the systematic distributional distance between representations of paired data from distinct modalities within a shared embedding space, despite the theoretical expectation that semantically identical pairs should align. Early empirical research first identified this geometric anomaly as a cone effect Liang et al. (2022), observing that embedding vectors tend to occupy a narrow cone Wang & Isola (2020). The C^3 framework Zhang et al. (2024) provided the first formal description of this structure, characterizing the gap as a superposition of a constant orthogonal displacement and random alignment noise. While these frameworks offer a foundational understanding, they predominantly rely on the assumption that the noise term is isotropic. Our work challenges this simplification, demonstrating that the residual noise exhibits high anisotropy. The absence of this theoretical perspective has directly limited recent explorations into training MLLMs using pure text. Unicorn Yu et al. (2025) pioneered the use of the modality gap to convert text representation into the pseudo-visual representation. However, due to their reliance on simple mean shifting, which implies an isotropic assumption, the synthesized representation fails to match the complex geometric shape of the real visual representation. This further highlights the necessity of constructing a novel training paradigm based on precise covariance alignment.