GEORX: A REFERENCE-BASED BENCHMARK FOR ATTRIBUTION DIAGNOSIS OF CROSS-MODAL EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-modal encoders map multimodal data into a shared embedding space for semantic alignment and retrieval. However, geometric deviations in this space can cause cosine-based rankings to diverge from semantic relevance, limiting retrieval accuracy. Existing benchmarks primarily measure end-to-end retrieval accuracy, offering limited support for attributing retrieval failures to geometric causes. To address this limitation, we introduce **GeoRx**, a reference-based benchmark for geometric attribution diagnosis in cross-modal retrieval. For each $\langle \text{encoder}, \text{workload} \rangle$ pair, GeoRx constructs a geometric reference model matched to the encoder's embedding dimension and the workload's corpus size. Five diagnostic metrics derived from candidate–corpus, candidate–candidate, and within-list relationships identify geometric deviations against this reference without requiring additional information. These relationships respectively characterize candidate occurrence frequency across queries, similarity among retrieved candidates, and variation in query–candidate scores. We validate diagnostic accuracy through controlled deviation injection. Guided by the diagnoses, we map correction methods from existing work to the identified deviations for individual and joint correction. Experiments demonstrate target detection rates of 98.93–100.0% and an overall hit rate of 32.35% across diagnosis–correction mappings. Code is available at `https://anonymous.4open.science/r/GeoRx-3CC3/`.

1 INTRODUCTION

Cross-modal encoders provide the foundation for semantic alignment and large-scale retrieval by mapping multimodal data, such as text and images, into a shared embedding space (Radford et al., 2021). Ideally, cosine similarity in the embedding space should reflect the semantic relevance between a query and a candidate. However, the representation spaces learned by encoders often exhibit different geometric deviation, such as **imbalanced global distributions**, **overcrowded local neighborhoods**, and **insufficient discrimination within candidate lists** (Dinu et al., 2014). These deviations can cause cosine-based rankings to deviate from true semantic relevance, thereby limiting the final accuracy of cross-modal retrieval.

Existing cross-modal benchmarks mainly evaluate the retrieval performance of encoders using end-to-end metrics such as Recall, NDCG, and MRR (Wei et al., 2024; Jiang et al., 2024b; Xu et al., 2026; Zhou et al., 2025a). These metrics can reflect the quality of retrieval results, but they provide limited information about which geometric deviation causes a retrieval failure. When different encoders or retrieval workloads exhibit similar decreases in Recall, the underlying representation spaces may contain substantially different structural problems. Meanwhile, retrieval performance is also affected by the corpus size, embedding dimension, and workload distribution. Therefore, directly comparing end-to-end metrics cannot sufficiently characterize the geometry of the embedding space itself. Cross-modal retrieval still lacks an evaluation tool that can systematically diagnose geometric deviations without relying on ground-truth or correction results.

Constructing such a diagnostic benchmark involves three major challenges. (C1) Retrieval failures result from the combined effects of multiple factors, and a single end-to-end metric cannot distinguish different geometric factors. (C2) Geometric metrics need to remain comparable across

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.

different embedding dimensions and corpus sizes. Otherwise, the observed differences may simply be caused by scale effects rather than genuine deviations in encoder representations. (C3) The diagnostic process cannot rely on performance changes caused by correction methods for circular validation. Otherwise, an effective general correction method may conceal the original deviation, while an unsuccessful correction may be incorrectly interpreted as evidence that the diagnosis was wrong. Therefore, diagnostic metrics need to be independently validated in a controlled environment with clearly defined geometric deviation ground-truth.

To address these problems, we introduce **GeoRx**, a reference-based benchmark for geometric attribution diagnosis in cross-modal retrieval. For each $\langle \text{encoder}, \text{workload} \rangle$ pair, GeoRx constructs an **reference model** matched to the encoder's embedding dimension and the workload's corpus size. We derive five diagnostic metrics from *candidate–corpus*, *candidate–candidate*, and *within-list* relationships. These relationships respectively characterize candidate occurrence frequency across queries, similarity among retrieved candidates, and variation in query–candidate scores. Comparing these metrics against their reference distributions identifies specific geometric deviations without requiring additional information of the encoder and workload. To validate each diagnostic accuracy, we inject the corresponding geometric deviation into the embedding space under controlled conditions and evaluate whether the diagnostic metric correctly detects it. Guided by the diagnoses, we map correction methods from existing works to the identified deviations for individual and joint correction. Rather than developing new correction methods, we introduce a benchmark to evaluate geometric deviations in existing cross-modal encoders and evaluate whether correction methods from existing work effectively mitigate these deviations.

The main contributions of this paper are as follows:

- We introduce **GeoRx**, a benchmark that establishes geometric reference models and five diagnostic metrics for cross-modal retrieval. GeoRx identifies geometric deviations by evaluating these metrics against references based on embedding dimension and corpus size, thereby enabling attribution diagnosis without requiring additional information of the encoder and workload.
- We validate the diagnostic metrics through **controlled injection** of their corresponding geometric deviations, individually and in combination. We then map diagnosed deviations to existing correction methods and evaluate whether the selected methods improve retrieval accuracy.
- Extensive experiments demonstrate that GeoRx achieves **target detection rates of 100.0% and 98.93%** under single- and dual-deviation injections, respectively. Existing correction methods achieve an overall **hit rate of 32.35%** across diagnosis–correction mappings, while joint correction remains sensitive to fusion weights. These findings highlight the lack of effective joint correction method guided by multiple diagnostic metrics.

2 Related Works

2.1 Cross-modal Encoder

We organize cross-modal encoders into three broad categories according to how their retrieval representations are constructed. These categories comprise **native cross-modal pretrained encoders**, **retrieval-adapted encoders**, and **encoders based on multimodal large language models (MLLMs)**. This grouping provides a practical basis for our experimental selection of three encoders from each category, covering different architectures and training paradigms.

Native cross-modal pre-trained models establish general-purpose vision–language representations. *CLIP* (Radford et al., 2021) and ALIGN (Jia et al., 2021) popularized large-scale contrastive learning with dual-tower architectures, while *SigLIP* (Zhai et al., 2023) modified the matching objective for more efficient alignment. BLIP (Li et al., 2022) and *BLIP-2* (Li et al., 2023a) introduced richer interactions between visual and language components. OpenCLIP (Cherti et al., 2023), LiT (Zhai et al., 2022), CoCa (Yu et al., 2022), Florence (Yuan et al., 2021), and EVA-CLIP (Sun et al., 2023) further explored variations in training data, objectives, and architectures. CLIPA-v2 (Li et al., 2023b), FILIP (Yao et al., 2021), FLIP (Li et al., 2023c), VirTex (Desai & Johnson, 2021),

and ConVIRT (Zhang et al., 2022) investigated fine-grained interaction, alignment, and training efficiency.

Retrieval-adapted encoders specialize pre-trained representations for downstream retrieval through fusion, fine-tuning, or retrieval-oriented training objectives. *CLIP-SF* (Wei et al., 2024) performs score-level fusion across modalities, *BLIP-FF* (Wei et al., 2024) uses feature-level fusion, and *OpenCLIP-FFT* (Cherti et al., 2023; Jiang et al., 2024b) adapts the embedding space through full-parameter fine-tuning. Other approaches, including the E5 family (Wang et al., 2022), MagicLens (Zhang et al., 2024a), and ALBEF (Li et al., 2021), incorporate instruction learning, hard negatives, or task-specific training strategies.

MLLM-based encoders derive retrieval representations from models that jointly process visual and textual inputs. *VLM2Vec* (Jiang et al., 2024b) and *E5-V* (Jiang et al., 2024a) use such models to construct vision–language embeddings. *GME* (Zhang et al., 2024b), LamRA (Liu et al., 2025), MegaPairs (Zhou et al., 2025b), and LLaVA-derived architectures (Liu et al., 2023) further explore multimodal representation learning and retrieval. Related developments include ColPali (Faysse et al., 2025), which uses multiple visual tokens for document retrieval, and NV-Embed (Lee et al., 2025), which studies embedding learning with large language models.

2.2 Cross-modal Retrieval Benchmarks

Existing cross-modal retrieval benchmarks can be broadly divided into **end-to-end retrieval benchmarks** and **semantic probing benchmarks**.

End-to-end retrieval benchmarks use metrics such as Recall@k, NDCG, and MRR to evaluate retrieval performance. M-BEIR (Wei et al., 2024), MMEB (Jiang et al., 2024b), MultiHaystack (Xu et al., 2026), and MR^2-Bench (Zhou et al., 2025a) provide standardized evaluation settings across different retrieval tasks and modalities. These benchmarks effectively measure retrieval performance, but do not explicitly characterize the geometric cause underlying the observed results.

Semantic probing benchmarks focus on finer-grained capabilities. Winoground (Thrush et al., 2022), VL-Checklist (Zhao et al., 2022), SugarCrepe (Hsieh et al., 2023), and CREPE (Ma et al., 2023) use counterfactual pairs or semantic perturbations to evaluate compositionality, attribute binding, actions, and spatial relations. While they reveal encoder sensitivity to specific semantic phenomena, they generally do not measure geometric deviations.

2.3 Geometric Deviations and Correction Methods

Embedding geometry can affect nearest-neighbor retrieval beyond individual query–candidate similarities. Prior studies have examined the effects of **hubness and density imbalance**, **local neighborhood structure**, and **graph connectivity** (Dinu et al., 2014; Radenović et al., 2018; Donoser & Bischof, 2013).

Existing correction methods exploit different sources of information to improve rankings. **Cross-domain Similarity Local Scaling(CSLS)** adjusts similarity scores using local density estimates to reduce nearest-neighbor density bias (Conneau et al., 2017). **α-weighted Query Expansion(α-QE)** refines query representations using highly ranked candidates (Radenović et al., 2018). **Diffusion** propagates query relevance over a candidate kNN graph (Donoser & Bischof, 2013; Iscen et al., 2017). **Listwise Cross-Encoder(L-CE)** and **Pointwise Cross-Encoder(P-CE)** jointly process query and candidate content to capture interactions beyond independently computed embeddings (Nogueira & Cho, 2019).

GeoRx builds on these lines of work but treats embedding-space geometry as an explicit diagnostic target rather than assuming a particular correction mechanism in advance. It diagnoses multiple geometric deviations independently of correction methods and subsequently evaluates whether the resulting diagnoses can provide a basis for selecting existing correction methods.

3 Benchmark

GeoRx comprises three modules, as shown in Figure 1. The **reference model** provides a geometric baseline, **geometric attribution diagnosis** identifies deviations, and **correction mapping** applies

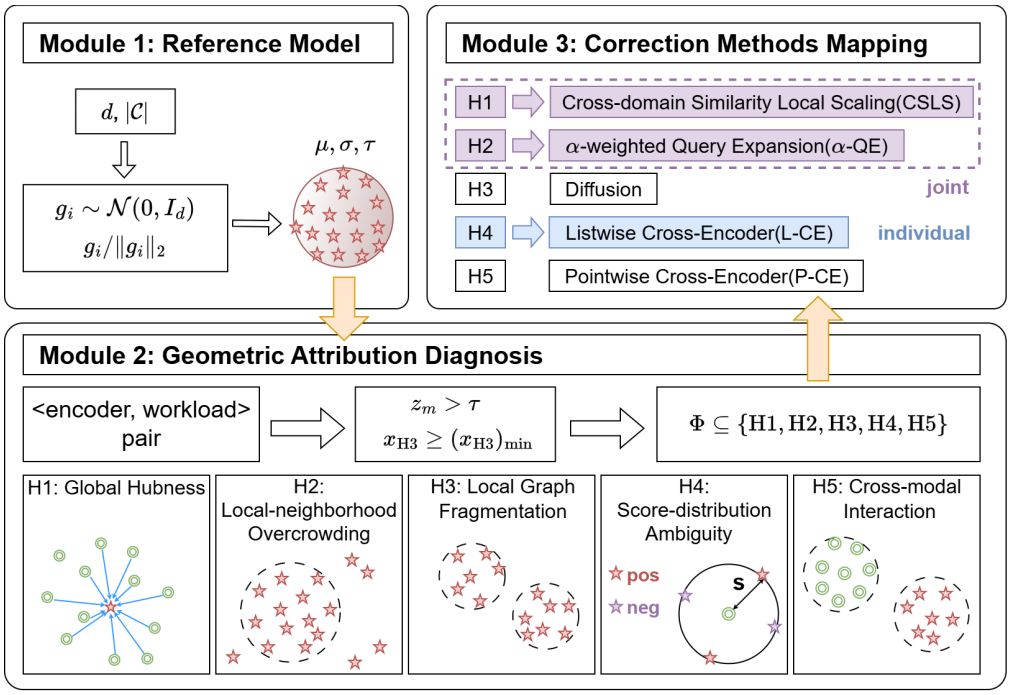

Figure 1: Architecture of GeoRx. The reference model supports geometric attribution diagnosis, which guides individual or joint correction methods to improve retrieval.

corresponding methods individually or jointly to improve Recall. For each $\langle \text{encoder}, \text{workload} \rangle$ pair, $Q_{diag}$is divided into three disjoint subsets: TRAIN for estimating metric centers and scales, CALIBRATION for determining the joint threshold, and FAULT_EVAL for original diagnosis and controlled injection validation. An independent evaluation set Q_{eval} is reserved for correction evaluation.

3.1 Reference Model

For a frozen encoder and a retrieval workload, let d be the embedding dimension and $|\mathcal{C}|$ the corpus size. Query and candidate embeddings $q, c \in \mathbb{R}^d$ are L2-normalized, and initial retrieval uses cosine similarity $s(q, c) = q^\top c$.

GeoRx uses an isotropic spherical distribution matched to d and $|\mathcal{C}|$ as the reference model for measuring geometric deviations.

Its vectors are generated by independently sampling $g_i \sim \mathcal{N}(0, I_d)$ and normalizing them as $\widetilde{g}_i = g_i/\|g_i\|_2$, where I_d is the identity matrix and $\widetilde{g}_i$ is a unit vector. This construction provides a reference null distribution without semantic structure. It does not represent an ideal retrieval space, and its applicability requires reassessment when the assumed normalization changes.

3.2 Geometric Attribution Diagnosis

GeoRx derives five diagnostic metrics without ground-truth from *candidate–corpus*, *candidate–candidate*, and *within-list relationships*. Let Q denote a query window and $L_n(q) = \text{Top-}n_{c \in \mathcal{C}} s(q, c)$ the top-n candidates for query q.

Candidate-corpus relationship, corresponds to the **global hubness metric H1**. For retrieval cutoff h, define occupancy count $N_h(c) = |\{q \in Q \mid c \in L_h(q)\}|$ (Radovanovic et al., 2010). Under a binomial opportunity model, each candidate enters a top-h list with probability $p = h/|\mathcal{C}|$, giving $N_h(c) \sim \text{Binomial}(|Q|, p)$. We normalize the observed occupancy skewness by its null value,

$$\mathrm{Skew}_{\mathrm{null}} = \frac{1-2p}{\sqrt{|Q|p(1-p)}}, \qquad x_{\mathrm{H1}} = \frac{G_1(N_h)}{\mathrm{Skew}_{\mathrm{null}}} \tag{1}$$

where $G_1(N_h)$ is the skewness of the observed occupancy counts with finite-sample bias correction. A larger x_{H1} indicates stronger global hubness.

Candidate-candidate relationships, include **local-neighborhood overcrowding metric H2** and **local graph fragmentation metric H3**. For H2, let $\mathrm{intra}(q)$ be the mean pairwise cosine similarity among distinct candidates in $L_n(q)$ and $\bar{s}_{\mathcal{C}}$ the corresponding background mean estimated from a randomly sampled candidate subset. We define

$$x_{\mathrm{H2}} = \frac{1}{|Q|} \sum_{q \in Q} \left(\mathrm{intra}(q) - \bar{s}_{\mathcal{C}}\right). \tag{2}$$

A larger x_{H2} indicate greater local concentration. The computation of $\mathrm{intra}(q)$ is detailed in Appendix A.1.

For H3, we construct a kNN graph on $L_n(q)$. Let $C_2(q)$ denote the size of the second-largest connected component, with $C_2(q) = 0$ when the graph is connected. The metric is

$$x_{\mathrm{H3}} = \frac{1}{|Q|} \sum_{q \in Q} \frac{C_2(q)}{n} \tag{3}$$

A larger x_{H3} indicates stronger local graph fragmentation.

Within-list relationships, include the **score-distribution ambiguity metric H4** and the **cross-modal interaction metric H5**. For H4, let $s_{(1)}(q)$ and $s_{(2)}(q)$ denote the cosine similarity of the first- and second-ranked candidates for query q. The distribution ambiguity metric is

$$x_{\mathrm{H4}} = \frac{1}{|Q|} \sum_{q \in Q} \left(s_{(1)}(q) - s_{(2)}(q)\right) \tag{4}$$

A smaller x_{H4} indicates weaker discrimination between the highest-ranked candidates.

For H5, we sample a candidate subset $A \subseteq \mathcal{C}$ and reuse it across all queries in Q. After removing each query's mean similarity, we decompose the remaining scores into a candidate effect shared across queries b_j and a query-specific interaction residual I_{ij}. Their energies are $E_{\mathrm{cand}} = \mathrm{mean}_j(b_j^2)$ and $E_{\mathrm{int}} = \mathrm{mean}_{i,j}(I_{ij}^2)$. Define the interaction energy ratio,

$$x_{\mathrm{H5}} = \frac{E_{\mathrm{int}}}{E_{\mathrm{int}} + E_{\mathrm{cand}}}. \tag{5}$$

We set $x_{\mathrm{H5}} = 0$ when both energies vanish. A smaller x_{H5} indicate weaker query-specific interaction. The full decomposition is provided in Appendix A.1.

Metric standardization and diagnostic results. To account for different metric scales and deviation directions, GeoRx estimates a center μ_m and a scale σ_m for each metric m from TRAIN windows. The standardized diagnostic score is

$$z_m = \begin{cases} \dfrac{x_m - \mu_m}{\sigma_m}, & m \in \{\mathrm{H1}, \mathrm{H2}, \mathrm{H3}\}, \\ \dfrac{\mu_m - x_m}{\sigma_m}, & m \in \{\mathrm{H4}, \mathrm{H5}\}. \end{cases} \tag{6}$$

Larger positive scores consistently indicate stronger deviations.

GeoRx calibrates a joint conformal threshold τ from the maximum standardized score across metrics in each CALIBRATION window, using a target family-wise false-positive level ε. H1, H2, H4, and

H5 are reported when $z_m > \tau$. H3 additionally requires $x_{\mathrm{H3}} \geq x_{\mathrm{H3,min}}$, where $x_{\mathrm{H3,min}}$ excludes small secondary fragments. The reported types form $\Phi \subseteq \{\mathrm{H1}, \mathrm{H2}, \mathrm{H3}, \mathrm{H4}, \mathrm{H5}\}$.

The complete procedures for estimating the centers and scales and calibrating the joint threshold are provided in Appendix A.2.

Controlled injection validates these diagnoses independently of correction outcomes. Predefined geometric deviations are applied only to a copy of FAULT_EVAL, while TRAIN and CALIBRATION observations remain unchanged.

3.3 Correction Methods

GeoRx freezes the diagnostic set Φ and maps its deviation types to predefined correction methods as shown in Figure 1, forming the predicted correction set P. The selected methods are applied individually or jointly to improve retrieval Recall.

On Q_{eval}, ground-truth is used to measure $\Delta\mathrm{Recall@}k$ relative to cosine retrieval for $k \in \{1, 5, 10\}$. Methods meeting the effectiveness criterion form the effective correction set T, which is compared with P to evaluate the mapping. Joint correction settings are evaluated separately.

4 Experiment

We evaluate GeoRx through geometric attribution diagnosis, retrieval correction, controlled injection, and two ablations. Implementation details and hardware configurations are provided in Appendix A.5.

4.1 Experiment Setup

We evaluate nine cross-modal encoders on ten M-BEIR workloads (Wei et al., 2024) as shown in Table 4.1, yielding 90 $\langle \mathrm{encoder}, \mathrm{workload} \rangle$ pairs. The encoders cover the three categories introduced in Related Works, with three models from each category. All experiments use precomputed embeddings without additional training or fine-tuning.

Encoder	Dim.	Workload	Candidates / Queries
E1: CLIP	512	W1: MSCOCO	5,000 / 24,809
E2: SigLIP	768	W2: VisualNews	542,246 / 19,995
E3: BLIP-2	256	W3: Fashion200K	201,824 / 1,719
E4: CLIP-SF	768	W4: WebQA	544,457 / 2,455
E5: BLIP-FF	768	W5: EDIS	1,047,067 / 3,241
E6: OpenCLIP-FFT	768	W6: NIGHTS	40,038 / 2,120
E7: GME	1,536	W7: OVEN	676,667 / 50,004
E8: VLM2Vec	3,072	W8: InfoSeek	611,651 / 11,323
E9: E5-V	4,096	W9: FashionIQ	74,381 / 6,003
		W10: CIRR	21,551 / 4,170

Table 1: Encoders and workloads. All nine encoders are evaluated on all ten workloads.

All embeddings are L2-normalized, and cosine retrieval is used as the baseline. For each $\langle \mathrm{encoder}, \mathrm{workload} \rangle$ pair, we repeat the experiments with three random seeds. Queries with valid positives in the corpus are sampled from the official test set and divided into disjoint subsets for diagnostic fitting, threshold calibration, diagnosis with controlled injection, and correction evaluation. Query counts and diagnostic list sizes are provided in Appendix A.5.

We evaluate CSLS, α-QE, diffusion, L-CE, and P-CE, together with selected combinations for joint correction. Retrieval performance is measured by Recall@k and its change from the cosine baseline $\Delta\mathrm{Recall@}k$, for $k \in \{1, 5, 10\}$. A correction is considered effective when $\Delta\mathrm{Recall@10} \geq 0.5\%$. Controlled injection is evaluated using the target detection rate, with predefined targets distinguished from successfully formed targets. Method configurations and detailed evaluation procedures are provided in Appendix A.5.

4.2 Geometric Attribution Diagnosis

GeoRx produces identical diagnostic sets across all three seeds for 78 of the 90 $\langle \text{encoder}, \text{workload} \rangle$ pairs. The remaining 12 pairs show agreement between two seeds. We take the majority diagnosis as the final set Φ, as shown in Figure 2. This agreement indicates stability across query samples.

The diagnoses vary with both encoder and workload. For CLIP, GeoRx identifies $\{\text{H1}, \text{H2}, \text{H5}\}$ on MSCOCO, WebQA, EDIS, FashionIQ, and CIRR. H1 is absent on VisualNews and NIGHTS, while H4 is additionally detected on Fashion200K, OVEN, and InfoSeek. In contrast, E5-V receives $\{\text{H1}, \text{H2}, \text{H5}\}$ on all ten workloads. CLIP-SF exhibits a different pattern, with H1 absent on MSCOCO and WebQA and H4 present on Fashion200K and NIGHTS. GeoRx thus captures both consistent encoder-level patterns and variation across workloads.

H3 is absent from all 90 pairs under the current graph construction and decision criterion. It requires the average relative size of the second-largest connected component to reach 15%, whereas the observed median and maximum are only 0.04% and 4.34% respectively. Although disconnected candidates occur, they form small fragments rather than substantial secondary components.

> **Insight 1.** ***Geometric deviations coexist and vary across encoders and workloads.***

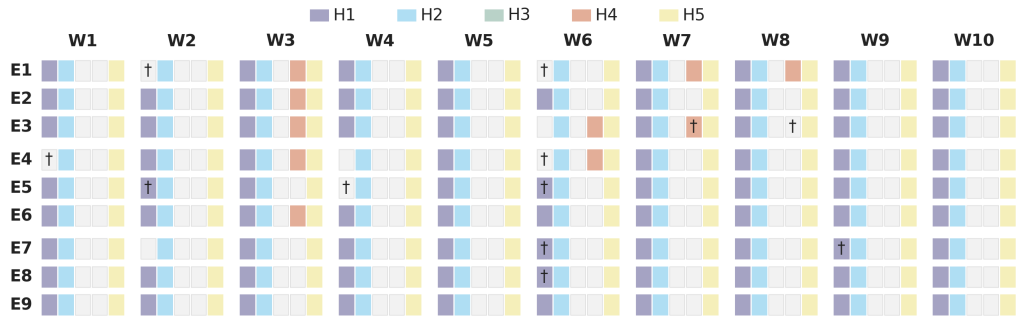

Figure 2: Geometric diagnoses across nine encoders and ten workloads. † marks a type on which one seed disagrees with the other two.

4.3 Correction Method Performance

Figure 3(a) summarizes the performance of five individual correction methods across all $\langle \text{encoder}, \text{workload} \rangle$ pairs. Mean $\Delta\text{Recall@10}$ is highest for CSLS at $+2.01\%$, followed by diffusion at -0.09%, α-QE at -0.30%, P-CE at -6.01%, and L-CE at -23.16%. CSLS and P-CE meet the effectiveness criterion in 64.07% and 34.44% of results, respectively, showing that a negative mean can mask improvements on some pairs.

For each diagnosis–method mapping, we measure the hit rate as the proportion of diagnosed pairs for which the corresponding correction method is effective. The overall hit rate is 32.35%. The specific hit rates are 64.63% for H1→CSLS, 3.33% for H2→ α-QE, 20.00% for H4→L-CE, and 33.33% for H5→P-CE. H3→diffusion cannot be assessed because no pair is diagnosed with H3.

Figure 3(b) compares individual and joint correction methods on selected pairs diagnosed with $\{\text{H2}, \text{H5}\}$. The joint correction methods combine α-QE and P-CE through weighted score fusion. Mean $\Delta\text{Recall@10}$ is -0.08% for α-QE alone and -3.84% for P-CE alone. It changes to $+0.56\%$ and -1.32% respectively for $w_{\text{QE}} = 0.75$ and 0.25, with $w_{\text{P-CE}} = 1 - w_{\text{QE}}$.

Fusion weights are chosen empirically and are not determined by the diagnosis, so applying these combinations to new settings requires tuning or a weight search. The same diagnostic set therefore does not guarantee that a fixed combination improves retrieval. These findings motivate joint correction methods that account for coexisting geometric deviations and their interactions.

> **Insight 2.** ***Correction methods do not necessarily improve retrieval under their corresponding geometric deviations, and the effectiveness of joint methods depends on the fusion weights.***

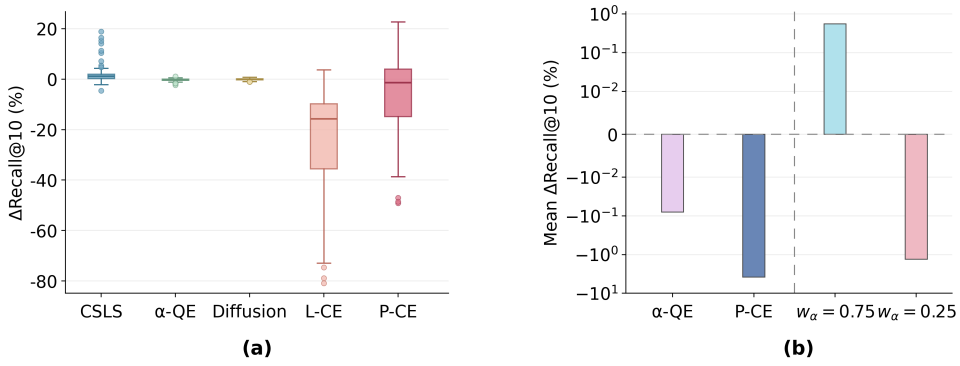

Figure 3: Correction performance measured by $\Delta\text{Recall@10}$ (%). (a) Individual methods across all $\langle \text{encoder}, \text{workload} \rangle$ pairs. (b) Individual methods and weighted fusion on selected pairs diagnosed with $\{\text{H2}, \text{H5}\}$.

4.4 Controlled Injection

Controlled injection evaluates GeoRx against predefined injection targets without relying on retrieval accuracy or correction outcomes. We consider single- and dual-deviation injections because additional simultaneous operations introduce interactions that make target formation harder to control. This restriction applies to the validation setting rather than the diagnostic scope of GeoRx.

We report the target detection rate, defined as the proportion of target deviations identified by GeoRx. This metric measures target coverage rather than exact agreement with the complete diagnostic set.

GeoRx achieves a target detection rate of 100.00% in single-deviation injections of H1–H5. For dual-deviation injections, the target detection rate is 98.93% against the predefined injection targets and reaches 100.00% when both targets are successfully formed. Figure 4 reports detection rates against the predefined targets. No successfully formed target is missed in either setting, supporting the sensitivity of GeoRx to individual deviations and their controlled pairwise coexistence.

> **Insight 3.** ***GeoRx effectively diagnoses target deviations under both single- and dual-deviation injections, supporting its ability to identify geometric deviations even when they coexist.***

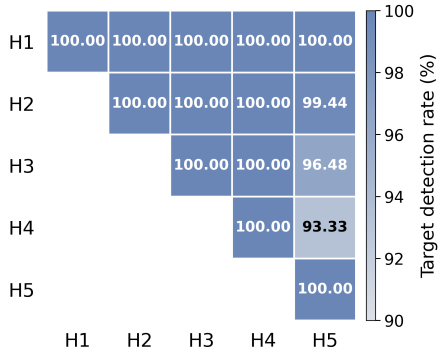

Figure 4: Target detection rates (%) under controlled injection. Diagonal and off-diagonal entries correspond to single- and dual-deviations respectively.

Configuration / Metric	Independent	Joint
$\langle 768, \text{MSCOCO} \rangle$	40.63	0.00
$\langle 768, \text{NIGHTS} \rangle$	31.25	3.13
$\langle 4096, \text{CIRR} \rangle$	25.00	0.00
Overall	**32.29**	**1.04**
H1	5.21	0.00
H2	9.38	0.00
H3	0.00	0.00
H4	8.33	1.04
H5	13.54	0.00

Table 2: False-positive rates (%) under independent and joint thresholding. Overall pools all settings. H1–H5 report per-metric rates.

4.5 Ablation on Reference Model Matching

We examine whether the reference model needs to match both d and $|\mathcal{C}|$ using six spherical configurations with $d \in \{768, 1536\}$ and $|\mathcal{C}| \in \{5000, 40038, 1047067\}$. Table 3 reports the reference centers μ_m and scales σ_m, both **multiplied by 10^3 for display**.

The reference center μ_m responds differently to d and $|\mathcal{C}|$. For H2, doubling d approximately halves μ_{H2} at every $|\mathcal{C}|$, whereas increasing $|\mathcal{C}|$ at $d = 768$ raises μ_{H2} from 0.00915 to 0.02203. For H4, μ_{H4} decreases with either d or $|\mathcal{C}|$ increase. It falls from 0.00920 to 0.00641 when d doubles at $|\mathcal{C}| = 5{,}000$, and to 0.00700 when the pool grows at $d = 768$. In contrast, the centers of H1 and H5 remain near 1 and 0.996 respectively, while the H3 center remains zero.

The reference scale σ_m can vary even when the center μ_m remains stable. At $d = 768$, increasing $|\mathcal{C}|$ from 5,000 to 1,047,067 reduces σ_{H1} from 0.0323 to 0.00173. The H4 and H5 scales also vary across spherical configurations, whereas σ_{H2} remains at its floor of 0.00100 and σ_{H3} remains at 0.02000.

> **Insight 4.** ***Reference metrics depend on both embedding dimension and corpus size, supporting a reference model matched to both for interpreting geometric deviations.***

d	$\lvert\mathcal{C}\rvert$	H1	H2	H4	H5
768	5,000	999.74 (32.25)	9.15 (1.00)	9.20 (0.66)	996.08 (0.41)
768	40,038	1000.00 (13.44)	14.06 (1.00)	7.98 (0.45)	996.14 (0.34)
768	1,047,067	999.86 (1.73)	22.03 (1.00)	7.00 (0.47)	996.03 (0.33)
1,536	5,000	996.65 (34.20)	4.65 (1.00)	6.41 (0.27)	996.09 (0.19)
1,536	40,038	1000.36 (12.93)	7.09 (1.00)	5.82 (0.26)	996.02 (0.32)
1,536	1,047,067	999.86 (2.59)	11.05 (1.00)	5.21 (0.35)	996.09 (0.22)

Table 3: Reference $\mathbf{10^3}\boldsymbol{\mu}$ $(\mathbf{10^3}\boldsymbol{\sigma})$ under different spherical configurations. The corresponding entry for H3 is 0.00×10^3 (20.00×10^3) in all configurations.

4.6 Ablation on Joint Conformal Thresholding

We compare independent per-metric thresholds (C-independent) with joint conformal thresholding (C-joint). The experiment uses spherical configuration $\langle 768, \mathrm{MSCOCO}\rangle$ with $(d, |\mathcal{C}|) = (768, 5000)$, $\langle 768, \mathrm{NIGHTS}\rangle$ with $(768, 40038)$, and $\langle 4096, \mathrm{CIRR}\rangle$ with $(4096, 21551)$. Each configuration contributes 32 held-out diagnostic windows excluded from TRAIN and CALIBRATION.

Both methods estimate reference centers and scales using the procedure from the main experiments. C-independent calibrates a separate threshold for each metric. C-joint calibrates a single threshold from the maximum standardized score across H1–H5.

Table 2 shows that C-joint reduces the overall window-level false-positive rate from 32.29% to 1.04%. Under C-joint, H1, H2, H3, and H5 produce no false positives, while H4 is reported in one window. These results support joint calibration for limiting false-positive accumulation when multiple diagnostic metrics are monitored.

> **Insight 5.** ***Joint conformal thresholding reduces false-positive rate compared with independent per-metric thresholds, substantially reducing false alarms when testing multiple deviations.***

5 Conclusion

In this paper, we presented **GeoRx**, a benchmark for geometric attribution diagnosis in cross-modal retrieval. GeoRx establishes reference models matched to embedding dimension and corpus size. By evaluating five diagnostic metrics against these references, GeoRx identifies which geometric deviations are present and whether they occur individually or coexist. Controlled injection experiments validate this capability, with target detection rates of 100.0% and 98.93% under single- and dual-deviation injections, respectively. This ability to identify individual and coexisting deviations enables systematic evaluation of whether existing correction methods are effective under each condition. Diagnosis-matched methods achieve an average hit rate of only 32.35%, while joint correction remains sensitive to fusion weights. By identifying coexisting deviations and revealing the limitations of existing methods in correcting them, GeoRx provides a basis for developing and evaluating joint correction guided by multiple diagnostic metrics.

AI USE STATEMENT

In this work, we used generative AI tools for to assist with translation. We have not used generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset support qualitative and thematic data analysis, or interpret results, and the formulation of mathematical claims, critical ingredients for proving mathematical claims, and assistance with writing proofs are not applicable to this work. We have reviewed all AI-assisted work. We checked translations against the intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The main text describes the reference model, diagnostic metrics, and evaluation protocol. Appendix A.1 provides the metric computations, and Appendix A.2 details standardization and threshold calibration. The controlled injection protocol and correction implementations are described in Appendices A.3 and A.4. Appendix A.5 summarizes the data partitions, experimental parameters, and computing environment.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. URL https://arxiv.org/abs/2107.07511.

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.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11157–11168. IEEE, 2021.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014.

Michael Donoser and Horst Bischof. Diffusion processes for retrieval revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1320–1327, 2013.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pp. 61424–61449, 2025.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, Teddy Furon, and Ondrej Chum. Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 926–935. IEEE, 2017.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024a.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024b.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. In *International Conference on Learning Representations*, volume 2025, pp. 79310–79333, 2025.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

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*, pp. 19730–19742. PmLR, 2023a.

Xianhang Li, Zeyu Wang, and Cihang Xie. An inverse scaling law for clip training. *Advances in Neural Information Processing Systems*, 36:49068–49087, 2023b.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23390–23400. IEEE, 2023c.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4015–4025. IEEE, 2025.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. @ crepe: Can vision-language foundation models reason compositionally? In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10910–10921. IEEE, 2023.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

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.

Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of machine learning research*, 11(sept):2487–2531, 2010.

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.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238. IEEE, 2022.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pp. 387–404. Springer, 2024.

Dannong Xu, Zhongyu Yang, Jun Chen, Yingfang Yuan, Ming Hu, Lei Sun, Luc Van Gool, Danda Pani Paudel, and Chun-Mei Feng. Multihaystack: Benchmarking multimodal retrieval and reasoning over 40k images, videos, and documents. *arXiv preprint arXiv:2603.05697*, 2026.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18102–18112. IEEE, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*, 2024a.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024b.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pp. 2–25. PMLR, 2022.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. An explainable toolbox for evaluating pre-trained vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 30–37, 2022.

Junjie Zhou, Ze Liu, Lei Xiong, Jin-Ge Yao, Yueze Wang, Shitao Xiao, Fenfen Lin, Miguel Hu Chen, Zhicheng Dou, Siqi Bao, et al. Mr 2-bench: Going beyond matching to reasoning in multimodal retrieval. *arXiv preprint arXiv:2509.26378*, 2025a.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. Megapairs: Massive data synthesis for universal multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19076–19095, 2025b.

A APPENDIX

A.1 METRIC COMPUTATION

All diagnostic metrics are computed from L2-normalized embeddings without using ground-truth. The corpus is kept fixed for each workload, and diagnostic retrieval uses cosine similarity over the full pool.

For H1, the occupancy count $N_h(c)$ includes every candidate in $\mathcal{C}$, including candidates with zero occurrences. The metric divides the finite-sample bias-corrected skewness of these counts by the binomial null skewness defined in the main text.

For H2, let $L_n(q) = \{v_1, \ldots, v_n\}$. The mean pairwise cosine similarity among distinct retrieved candidates can be computed without explicitly forming all pairs,

$$\mathrm{intra}(q) = \frac{\left\| \sum_{j=1}^{n} v_j \right\|_2^2 - \sum_{j=1}^{n} \|v_j\|_2^2}{n(n-1)}. \tag{7}$$

Because the embedding vectors are normalized, this expression equals the average of $v_i^\mathsf{T} v_j$ over $i \neq j$. The background mean $\bar{s}_\mathcal{C}$ is estimated from a randomly sampled candidate subset using the same pairwise computation. H2 averages $\mathrm{intra}(q) - \bar{s}_\mathcal{C}$ over the query window Q.

For H3, we construct a kNN graph among the candidates in $L_n(q)$, excluding self-neighbors. The graph is made undirected by retaining an edge whenever either endpoint selects the other as a neighbor. Connected components are ordered by size, and the size of the second-largest component is divided by the diagnostic list size $n = |L_n(q)|$. This fraction is zero for a connected graph and is averaged over queries to obtain x_{H3}. The additional requirement $x_{\mathrm{H3}} \geq x_{\mathrm{H3,min}}$ applies to this window-level average.

For H4, candidates are ordered by cosine similarity. The metric is the mean difference between the first- and second-ranked scores across queries.

For H5, we sample a candidate subset $A \subseteq \mathcal{C}$ uniformly without replacement, with $|A| = \min(1024, |\mathcal{C}|)$. Sampling uses NumPy with a fixed seed of 42. Within each $\langle$encoder, workload$\rangle$ pair, the subset A is used across TRAIN, CALIBRATION, and FAULT_EVAL windows. The score matrix is computed between the queries in each window and the candidates in A. Let $q_i \in Q$, $a_j \in A$, and $S_{ij} = q_i^\mathsf{T} a_j$. We first remove each query's mean similarity,

$$R_{ij} = S_{ij} - \frac{1}{|A|} \sum_{j'=1}^{|A|} S_{ij'}. \tag{8}$$

The centered scores are decomposed into a candidate effect shared across queries and a query-specific interaction residual,

$$b_j = \frac{1}{|Q|} \sum_{i=1}^{|Q|} R_{ij}, \qquad I_{ij} = R_{ij} - b_j. \tag{9}$$

Their energies are

$$E_{\mathrm{cand}} = \frac{1}{|A|} \sum_j b_j^2, \qquad E_{\mathrm{int}} = \frac{1}{|Q||A|} \sum_{i,j} I_{ij}^2. \tag{10}$$

The resulting metric is

$$x_{\mathrm{H5}} = \frac{E_{\mathrm{int}}}{E_{\mathrm{int}} + E_{\mathrm{cand}}}. \tag{11}$$

When both energies vanish, the metric is set to zero. Otherwise, it lies in $[0, 1]$. Smaller values indicate a larger contribution from shared candidate preferences relative to query-specific interaction.

A.2 CALIBRATION PROCEDURE

For the main diagnosis, the reference centers μ_m, scales σ_m, and threshold τ are estimated from the matched spherical reference model. The self-reference comparison instead estimates these parameters from the observed embedding space's TRAIN and CALIBRATION subsets. Parameters from the self-reference comparison are not used to produce the main diagnostic set Φ.

For each monitored statistic m, let $x_{m,j}$ be its value on the j-th of r TRAIN windows. The reference center is

$$\mu_m = \mathrm{median}_{1 \le j \le r}\, x_{m,j}. \tag{12}$$

We estimate dispersion using the median absolute deviation and interquartile range,

$$\begin{aligned} \mathrm{MAD}_m &= \mathrm{median}_j\, |x_{m,j} - \mu_m|, \\ \mathrm{IQR}_m &= \mathrm{Quantile}_{0.75}(x_m) - \mathrm{Quantile}_{0.25}(x_m). \end{aligned} \tag{13}$$

The scale is

$$\sigma_m = \max\left(c_{\mathrm{MAD}}\, \mathrm{MAD}_m,\ \frac{\mathrm{IQR}_m}{c_{\mathrm{IQR}}},\ \mathrm{floor}_m \right), \tag{14}$$

where $c_{\mathrm{MAD}} \approx 1.4826$ and $c_{\mathrm{IQR}} \approx 1.349$ are normal-consistency constants. The positive scale floor prevents nearly constant statistics from producing unstable standardized scores.

Scores are oriented so that larger positive values indicate stronger deviations,

$$z_m = \begin{cases} (x_m - \mu_m)/\sigma_m, & \text{for an upper-tail deviation,} \\ (\mu_m - x_m)/\sigma_m, & \text{for a lower-tail deviation.} \end{cases} \tag{15}$$

H1, H2, and H3 use the upper-tail direction, while H4 and H5 use the lower-tail direction.

The stored calibration profiles also monitor two auxiliary statistics, the occupancy coefficient of variation normalized by its binomial null value and the mean local clustering coefficient. The former uses the upper-tail direction and the latter the lower-tail direction. Thus, the calibrated family contains seven statistics, while the reported geometric deviation types remain H1–H5.

Let $\mathcal{M}$ denote the monitored family. For CALIBRATION window i, we compute

$$U_i = \max\left(0, \max_{m \in \mathcal{M}} z_{m,i}\right). \tag{16}$$

After sorting the n_{cal} values in ascending order, we set

$$\ell = \lceil (n_{\mathrm{cal}} + 1)(1 - \varepsilon) \rceil, \qquad \tau = \max\left(U_{(\ell)}, 10^{-8}\right). \tag{17}$$

The spherical profiles use $n_{\mathrm{cal}} = 32$ and $\varepsilon = 0.05$, giving $\ell = 32$. The threshold is therefore the largest calibration score, subject to the numerical lower bound.

Let U_{new} denote the maximum directional standardized score in a new evaluation window drawn from the reference null distribution excluded from Q_{eval}. Then $\Pr(U_{\mathrm{new}} > \tau)$ is the probability that at least one monitored statistic exceeds the joint threshold. Under exchangeability of CALIBRATION and new evaluation windows from the reference null distribution, this order statistic provides the split-conformal bound (Angelopoulos & Bates, 2022)

$$\Pr(U_{\mathrm{new}} > \tau) \le \frac{n_{\mathrm{cal}} + 1 - \ell}{n_{\mathrm{cal}} + 1}. \tag{18}$$

For the stated configuration, the bound is $1/33$. This interpretation concerns reference-null windows and does not imply a false-positive guarantee for arbitrary real embedding distributions.

The reference center, scale, and threshold remain fixed during diagnosis and controlled injection. The threshold ablation evaluates additional clean spherical evaluation windows that are excluded from fitting and calibration.

A.3 Controlled Injection

Controlled injection evaluates whether GeoRx identifies predefined injection targets independently of retrieval correction. We consider each of the five geometric deviation types individually and all ten pairs of types. Injections are applied to copies of the evaluation observations. The observations used for fitting and calibration remain unchanged, and no diagnostic parameter is refitted after injection.

The intended geometric effects correspond to the five metric definitions. H1 targets concentration of candidate occupancy across queries. H2 targets increased similarity among retrieved candidates. H3 targets the formation of substantial secondary graph components. H4 targets a reduced margin between the highest-ranked candidates.

The H5 injection contracts queries toward a common direction estimated from the clean training queries. Let Q_{train} denote this query set and define

$$\bar{q} = \frac{\sum_{q \in Q_{\mathrm{train}}} q}{\left\| \sum_{q \in Q_{\mathrm{train}}} q \right\|_2}.$$

Each query is transformed as

$$q_i' = \frac{(1-t)q_i + t\bar{q}}{\|(1-t)q_i + t\bar{q}\|_2}, \qquad t = 0.75.$$

The corpus remains unchanged. The transformation uses neither diagnostic outputs nor calibrated thresholds or relevance labels.

For controlled joint injections into real embedding spaces, all target-specific operators use a shared strength of 0.75. A target deviation is considered successfully formed when at least 90% of queries satisfy its geometric acceptance condition. These conditions assess increased candidate-occurrence concentration and supporting-query coverage for H1, increased mean pairwise similarity within the top-50 list for H2, a nonzero second-largest connected component for H3, a top-1–top-2 score gap of at most 10^{-5} for H4, and a reduced interaction-energy share for H5. The H1, H2, and H5 conditions are evaluated relative to the clean embeddings. The coverage requirement concerns queries rather than embedding coordinates.

We distinguish predefined targets from targets that satisfy the geometric formation criteria. For evaluation case t, let G_t be the predefined target set, $F_t \subseteq G_t$ the successfully formed target set, and Φ_t the diagnostic output. The target detection rate is

$$\mathrm{TDR} = \frac{\sum_t |\Phi_t \cap G_t|}{\sum_t |G_t|}. \tag{19}$$

The geometric formation rate and detection rate conditional on target formation are

$$\mathrm{GFR} = \frac{\sum_t |F_t|}{\sum_t |G_t|}, \qquad \mathrm{TDR}_{\mathrm{formed}} = \frac{\sum_t |\Phi_t \cap F_t|}{\sum_t |F_t|}. \tag{20}$$

The conditional rate excludes unformed targets from its denominator. It does not require every target in a case to have formed.

All predefined targets contribute to the unconditional detection rate, including those that fail the formation criteria. Neither retrieval Recall nor correction outcomes enter these diagnostic evaluation measures.

A.4 Correction Implementation

The correction mapping associates H1–H5 with CSLS, α-QE, diffusion, L-CE, and P-CE respectively. Individual methods are evaluated relative to the same cosine baseline on Q_{eval}. The diagnostic set is fixed before correction evaluation.

CSLS applies a candidate density penalty. (Conneau et al., 2017) Let Q_{CSLS} be the reference query set and $\mathcal{N}_{n_{\mathrm{CSLS}}}(c; Q_{\mathrm{CSLS}})$ the queries with the highest cosine similarity to candidate c. We compute

$$r_{\mathcal{C}}(c) = \frac{1}{n_{\mathrm{CSLS}}} \sum_{q' \in \mathcal{N}_{n_{\mathrm{CSLS}}}(c; Q_{\mathrm{CSLS}})} s(q', c), \tag{21}$$

and rank candidates using

$$s_{\mathrm{CSLS}}(q, c) = s(q, c) - \lambda r_{\mathcal{C}}(c). \tag{22}$$

The reference queries are drawn from the diagnostic TRAIN subset. A penalty depending only on the current query is constant across candidates and can be omitted without changing their order.

For α-QE, let $c_{(j)}$ be the j-th candidate under cosine retrieval. (Radenović et al., 2018) The neighbor weights are

$$a_j = \frac{[s(q, c_{(j)})]_+^{p_{\mathrm{QE}}}}{\sum_{\ell=1}^{n_{\mathrm{QE}}} [s(q, c_{(\ell)})]_+^{p_{\mathrm{QE}}} + 10^{-8}}, \tag{23}$$

where $[x]_+ = \max(x, 0)$. The expanded query is

$$q' = \mathrm{normalize}\left((1-\alpha)q + \alpha \sum_{j=1}^{n_{\mathrm{QE}}} a_j c_{(j)} \right). \tag{24}$$

Candidates are then ranked by $q'^{\mathsf{T}} c$. The mixing coefficient α is distinct from the weighting exponent p_{QE} and the joint correction fusion weight.

Diffusion uses a directed candidate kNN graph. (Donoser & Bischof, 2013; Iscen et al., 2017) For candidate c_i, let $\mathcal{N}_{10}(i)$ contain its 10 nearest candidates by cosine similarity, excluding itself. The propagation matrix $W \in \mathbb{R}^{|\mathcal{C}| \times |\mathcal{C}|}$ is defined by

$$W_{ij} = \begin{cases} \dfrac{\max(c_i^{\mathsf{T}} c_j, 0)}{\sum_{\ell \in \mathcal{N}_{10}(i)} \max(c_i^{\mathsf{T}} c_\ell, 0) + 10^{-8}}, & j \in \mathcal{N}_{10}(i), \\ 0, & \text{otherwise.} \end{cases}$$

Thus, nonnegative cosine weights are normalized by row, without additional self-loops or graph symmetrization. For each query, one propagation step updates the cosine similarity vector s_0 as

$$s_1 = (1-\beta)s_0 + \beta W s_0, \qquad \beta = 0.10.$$

Candidates are ranked by the resulting scores.

Both P-CE and L-CE rerank the top 100 candidates retrieved by cosine similarity. Their implementations depend on the workload and use frozen BLIP features or pretrained BLIP matching scores.

For MSCOCO, VisualNews, Fashion200K, WebQA, EDIS, OVEN, and InfoSeek, P-CE uses the pretrained `Salesforce/blip-itm-base-coco` checkpoint without local training and ranks candidates by its image–text matching score. For NIGHTS, FashionIQ, and CIRR, P-CE uses a locally trained head over frozen BLIP vision features. (Li et al., 2022) Let $u_q, u_c \in \mathbb{R}^D$ denote the query and candidate features. The head applies a linear projection from $2D$ to D, GELU, and a linear projection to a scalar score to their concatenation $[u_q; u_c]$. Training minimizes pointwise binary cross-entropy plus 0.5 times a listwise cross-entropy term computed over one positive and seven negatives. We use AdamW with a learning rate of 10^{-3} and a batch size of 32 for at most 4,000 steps. Validation is performed every 200 steps. Training stops after eight consecutive evaluations without improvement in validation binary cross-entropy, and the checkpoint with the lowest validation loss is retained.

L-CE uses a locally trained ListMixer over frozen BLIP image and text features. (Li et al., 2022) Each concatenated query–candidate feature is projected from $2D$ to D, followed by GELU and LayerNorm. The resulting candidate representations are processed by a two-layer Transformer encoder with eight attention heads, a feed-forward dimension of $4D$, and dropout of 0.1. L-CE outputs one score per candidate. Each training list contains one positive and 99 negatives, matching the reranking depth at inference. Candidate order is randomized, and the positive index is tracked explicitly. Each workload provides 4,096 training lists and 1,024 validation lists. We use AdamW with a learning rate of 3×10^{-4}, weight decay of 10^{-2}, a batch size of 16, and gradient clipping at 1.0 for at most 6,000 steps. Validation is performed every 200 steps. Training stops after eight consecutive evaluations without an improvement of at least 10^{-4} in validation cross-entropy. The checkpoint with the lowest validation cross-entropy is retained.

The joint experiment for $\{\mathrm{H2}, \mathrm{H5}\}$ combines α-QE and P-CE on a shared cosine top-100 candidate list. The weights assigned to α-QE are $w_{\mathrm{QE}} \in \{0.25, 0.50, 0.75\}$, with complementary P-CE

weights $1 - w_{\mathrm{QE}}$. These are empirical configurations rather than weights determined by the diagnostic set. The reported subset contains four distinct $\langle \text{encoder}, \text{workload} \rangle$ pairs and ten pair–seed observations.

For a query q, let $\mathcal{R}(q)$ be its relevant candidates and $L_k^{(a)}(q)$ the top-k list under method a. The implementation evaluates query-level retrieval success as

$$\mathrm{Recall@}k(a) = \frac{1}{|Q_{\mathrm{eval}}|} \sum_{q \in Q_{\mathrm{eval}}} \mathbf{1}\{L_k^{(a)}(q) \cap \mathcal{R}(q) \neq \}. \tag{25}$$

The gain is

$$\Delta\mathrm{Recall@}k = \mathrm{Recall@}k(a) - \mathrm{Recall@}k(\mathrm{cosine}). \tag{26}$$

A correction is effective when $\Delta\mathrm{Recall@}10 \geq 0.5\%$. Diagnosis-conditioned effectiveness is evaluated over pairs containing the corresponding diagnosed type, with effectiveness determined by a majority of seeds.

A.5 Experiment Setup

For each $\langle \text{encoder}, \text{workload} \rangle$ pair, queries with at least one relevant candidate in the corpus are selected from the official test queries. Each seed generates a new partition containing 512 TRAIN queries, 320 CALIBRATION queries, 320 FAULT_EVAL queries, and 500 correction evaluation queries. These subsets are disjoint within a seed, while partitions from different seeds may overlap. The main experiments use seeds 1, 2, and 3.

The observed-space TRAIN and CALIBRATION subsets support the self-reference comparison. The main spherical reference parameters are fitted from separately generated spherical observations. The spherical profiles use 32 TRAIN windows and 32 CALIBRATION windows, each containing 256 queries. Observed-space diagnosis uses two windows from FAULT_EVAL. The local diagnostic implementation aggregates their raw metric values by the median before applying the reference parameters.

Table 4: Diagnostic and correction parameters.

Parameter	Value
Diagnostic window size	256
Diagnostic list size n	50
H1 occupancy cutoff h	10
H3 graph neighbor count	10
H3 minimum component fraction	0.15
Spherical TRAIN windows	32
Spherical CALIBRATION windows	32
Target false-positive level ε	0.05
CSLS neighbor count n_{CSLS}	10
CSLS penalty coefficient λ	0.50
Query expansion neighbor count n_{QE}	10
Query mixing coefficient α	0.10
Query expansion exponent p_{QE}	3
Diffusion graph neighbor count	10
Diffusion coefficient β	0.10
Diffusion steps	1
Cross-encoder reranking cutoff N	100
H5 candidate subset size	$\min(1024, \lvert\mathcal{C}\rvert)$
H5 candidate sampling seed	42
Joint injection strength	0.75
Geometric formation coverage	0.90
H4 formation score-gap tolerance	10^{-5}
Diffusion neighbors	10
Diffusion propagation steps	1
Diffusion mixing coefficient β	0.10
CE reranking depth	100
L-CE training list size	100

The scale floors for H1–H5 are 10^{-3}, 10^{-3}, 0.02, 10^{-4}, and 10^{-5} respectively. The auxiliary occupancy and clustering statistics use floors of 10^{-3} and 0.02. In the reference-matching table, all six H2 scales equal the floor of 10^{-3}. This equality does not imply identical unregularized dispersion across configurations.

For the joint-threshold ablation, each of the three spherical configurations contributes 32 spherical evaluation windows, giving 96 windows per protocol. These spherical evaluation windows are separate from the TRAIN and CALIBRATION windows. The additional real-space reference and threshold comparisons use two seeds.

Experiments use Python 3.10 and PyTorch 2.6.0 on Ubuntu 20.04.6, with two Intel Xeon Gold 5218R CPUs, 512 GiB RAM, and four 32 GiB NVIDIA Tesla V100 GPUs. Some embeddings are precomputed on NVIDIA A100 GPUs.