Generative Giants, Retrieval Weaklings: Why do Multimodal Large Language Models Fail at Multimodal Retrieval?

Hengyi Feng[1,2], **Zeang Sheng**[2], **Meiyi Qiang**[2], **Yang Li**[3], **Wentao Zhang**[2,4†]
[1] University of Electronic Science and Technology of China
[2] Peking University [3] Tencent Inc [4] Zhongguancun Academy
hengyi.feng@std.uestc.edu.cn, *wentao.zhang@pku.edu.cn*

Abstract

Despite the remarkable success of multimodal large language models (MLLMs) in generative tasks, we observe that they exhibit a counterintuitive deficiency in the zero-shot multimodal retrieval task. In this work, we investigate the underlying mechanisms that hinder MLLMs from being effective retrievers. With the help of sparse autoencoders (SAEs), we decompose MLLM output representations into interpretable semantic concepts to probe their intrinsic behavior. Our analysis reveals that the representation space of MLLMs is overwhelmingly dominated by textual semantics; and the visual semantics essential for multimodal retrieval only constitute a small portion. We find that this imbalance is compounded by the heavy focus of MLLMs on bridging image-text modalities, which facilitates generation but homogenizes embeddings and finally diminishes the discriminative power required for multimodal retrieval. We further discover that the specific feature components that contribute most to the similarity computations of MLLMs are actually distractors that greatly reduce retrieval performance. Building on these insights, we propose ReAlign, a test-time adaptation approach that applies a whitening transformation to adjust the geometry of MLLM representation spaces. Empirical results show that this simple intervention consistently improves zero-shot multimodal retrieval performance across diverse MLLMs without fine-tuning efforts. The code is available at `https://github.com/Heinz217/mllm-retrieval-analysis`.

1 Introduction

Rapid advances in multimodal large language models (MLLMs) (Xu et al., 2025; Team et al., 2025; Steiner et al., 2024; Liu et al., 2024a) have revolutionized cross-modality understanding, establishing state-of-the-art performance in generative tasks such as image captioning, visual question answering (VQA), and complex visual reasoning (Caffagni et al., 2024; Bai et al., 2024). This proficiency mainly stems from MLLM's massive parameter scale (Kaplan et al., 2020) that enables the storage of rich parametric knowledge (Cheng et al., 2024) and the capture of deep semantic dependencies (Zhang et al., 2025a). Consequently, there is a growing interest in exploring the potential of MLLMs for tasks extending pure generative applications (Jiang et al., 2024b; Cai et al., 2025).

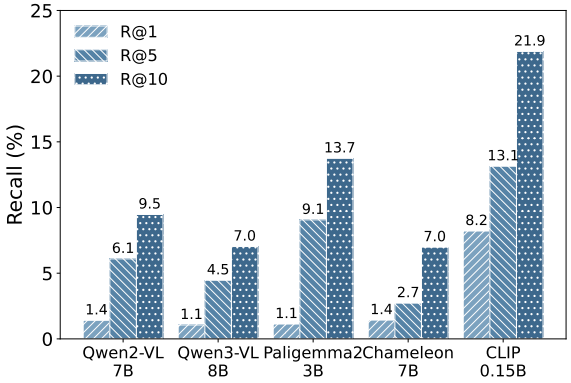

Figure 1: Multimodal retrieval performance of MLLMs and CLIP on the CIRR $((q_i, q_t) \rightarrow c_i)$ dataset. The results illustrate the inferior performance of MLLMs.

As a core task in information retrieval (IR), multimodal retrieval aims to locate relevant content across different data modalities. A dominant paradigm within this field is dense retrieval, where queries and candidates (e.g., images and texts) are mapped into a shared high-dimensional space. Under this setting, the semantic relationship between modalities is measured by embedding similarity (Fang et al., 2024). The models responsible for encoding inputs into vector representations are typically referred to as retrievers.

Intuitively, the extensive parametric memory and the excellent semantic understanding of MLLMs should equip them to act as capable retrievers.

[†] Corresponding Author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 15917–15933

However, in practice, we observe a counterintuitive phenomenon: embeddings directly derived from MLLMs yield poor performance in the zero-shot multimodal retrieval task (Figure. 1). Surprisingly, conventional contrastive vision-language models (e.g., CLIP (Radford et al., 2021)) even significantly outperform MLLMs, although MLLMs have far more parameters than the former. While MLLMs are not explicitly optimized for representation learning objectives, the magnitude of this performance gap suggests a fundamental limitation when repurposing these generative giants as multimodal retrievers.

Identifying the limitations behind this failure may offer key insights for expanding the use of MLLMs in multimodal retrieval. In this paper, we aim to answer the following research question:

What hinders MLLMs from being effective multimodal retrievers?

To address this question, we dissect the internal representational mechanisms of MLLMs. We employ sparse autoencoders (SAEs) (Huben et al., 2024) as the key analytical tool to disentangle dense representations into linear combinations of interpretable semantic concepts. Drawing inspiration from recent work (Papadimitriou et al., 2025), we introduce quantitative metrics to probe the representational space from multiple angles and explicitly link these properties to their impact on the multimodal retrieval performance of MLLMs. Rigorous empirical experiments that train SAEs on top of billions of activations are conducted on a diverse set of MLLM architectures (Bai et al., 2025; Wang et al., 2024; Team, 2024; Steiner et al., 2024).

Through these analyses, we identify three primary factors that limit MLLMs from functioning as effective multimodal retrievers:

- We demonstrate that the representational spaces of MLLMs are dominated by the text modality. This inherent bias constrains the capacity to encode visual information essential for retrieval.
- We analyze how MLLMs allocate their representational budget and find a heavy concentration on bridging the modality gap. While beneficial for generation, this alignment reduces the distinctiveness of the derived embeddings.
- When acting as retrievers, the dominant components on MLLMs's similarity computation turn out to be counterproductive distractors for multimodal retrieval.

Guided by these insights, we propose ReAlign, a Test-Time Adaptation intervention designed to rectify the geometry of MLLM feature spaces. By applying a whitening transformation, ReAlign mitigates representational bottlenecks without requiring additional fine-tuning or labeled data. Extensive experiments demonstrate that our method unlocks the latent retrieval capabilities of MLLMs, consistently yielding significant retrieval performance improvements across diverse architectures.

In summary, the main contributions and benefits of our work are as follows:

- To the best of our knowledge, our work is the first study to analyze MLLM representations and identify intrinsic factors that compromise their performance on multimodal retrieval tasks.
- We propose ReAlign, a training-free solution that realigns the embedding spaces of MLLMs, enabling off-the-shelf retrieval performance improvements for MLLMs without finetuning.

2 Preliminary: Zero-shot multimodal retrieval with MLLMs

In this work, we study the multimodal retrieval task in the zero-shot setting. Given a query q, the goal is to retrieve a list of candidates $\{c_1, c_2, \ldots, c_k\} \subset \mathcal{C}$ to maximize ranking metrics such as Recall@1. The queries and candidates can be text, image, or mixed text–image inputs: $q \in \{q_t, q_i, (q_t, q_i)\}$; $c \in \{c_t, c_i, (c_t, c_i)\}$. We use the hidden states of multimodal large language models (MLLMs) to extract embeddings for both queries and candidates.

Although MLLMs can produce hidden state representations, they are not originally designed as embedding models. Previous works (Liu et al., 2025; Lin et al., 2025) typically use the hidden state of the last token (or special tokens) as the embedding, but such representations work well only when the model has been explicitly optimized for representation learning. In the zero-shot setting considered here, where no such optimization is allowed, these embeddings are suboptimal. Instead, we observe that mean pooling over all hidden states yields a consistently better performance (see Appendix A for details). In this work, unless otherwise specified, we adopt mean pooling over the hidden states of the last layer as the embedding for all MLLM-based multimodal retrieval experiments.

3 What hinders MLLMs from being effective retrievers? A mechanistic analysis with sparse autoencoders

To analyze why MLLMs underperform in zero-shot multimodal retrieval scenarios, we perform a mechanistic interpretability analysis using sparse autoencoders (SAEs) (Olshausen and Field, 1997; Makhzani and Frey, 2014).

SAEs have gained widespread popularity in recent years due to their remarkable ability to interpret language model activations (Bricken et al., 2023; Rajamanoharan et al., 2024). By reconstructing internal representations with sparsely activated features, SAEs disentangle them into semantic concepts (Yin et al., 2025; Kissane et al., 2024; Makelov et al., 2024). Given the latent representations $H \in \mathbb{R}^{n \times d}$, the corresponding sparse codes $Z \in \mathbb{R}^{n \times c}$ are computed as:

$$Z = \Phi(HW_{enc}^{\top} + b), \quad \hat{H} = ZD, \tag{1}$$

where $W_{enc} \in \mathbb{R}^{c \times d}$ is the learned weight matrix, $D \in \mathbb{R}^{c \times d}$ is the learned dictionary, $b \in \mathbb{R}^{c}$ is the bias vector, and $\Phi(\cdot)$ denotes a nonlinear activation function such as ReLU. Specifically, each row of D can be regarded as a distinct concept vector, capturing an interpretable direction. The reconstruction loss of the SAE is then defined as:

$$\mathcal{L}_{rec}(H) = \|H - \hat{H}\|^2 + \alpha\|Z\|_1, \tag{2}$$

where the first term ensures faithful reconstruction of the input representation, and the second term enforces sparsity on the latent code through the ℓ_1-norm, controlled by a parameter α.

In our work, we consider representative MLLMs with different architectures, including Qwen3-VL-8B-Instruct (Bai et al., 2025), Qwen2-VL-7B-Instruct (Wang et al., 2024) , Chameleon-7B (Team, 2024), and Paligemma2-3B-Mix-224 (Steiner et al., 2024). For contrastive vision language models (VLMs), we consider CLIP (Clip-ViT-Base-Patch32) (Radford et al., 2021) and SigLIP2 (SigLIP2-Base-Patch16-512) (Tschannen et al., 2025). In our experiments, we employ Top-K SAEs (Gao et al., 2024) to learn interpretable concept representations from the activations of the COCO (Lin et al., 2014) dataset, using the last-layer hidden states for MLLMs and the last-layer embeddings for contrastive VLMs. More implementation details are provided in Appendix B.

3.1 Evaluation metrics of learned concepts

To evaluate and analyze the learned representations, we introduce four distinct metrics: ***energy***, ***modality score***, ***bridge score*** (Papadimitriou et al., 2025), and ***retrieval attribution score***. Each measure provides unique insights by focusing on a specific dimension of the representational space.

Energy. Energy represents how frequently and strongly a concept is activated across samples. Specifically, for concept i, it is defined as the expected activation magnitude over all samples:

$$\text{Energy}_i = \mathbb{E}_z \left[z_i \right]. \tag{3}$$

Concepts with higher energy are activated more frequently or strongly, capturing more dominant or widely shared patterns in the representation space.

Modality Score. The modality score quantifies a concept's bias towards text or image. For concept i, the modality score is computed as follows:

$$\text{ModalityScore}_i = \frac{\mathbb{E}_{z \sim \tau} \left[z_i \right]}{\mathbb{E}_{z \sim \iota} \left[z_i \right] + \mathbb{E}_{z \sim \tau} \left[z_i \right]}, \tag{4}$$

where ι and τ denote the image and text modalities, respectively. For a given concept, higher scores indicate text dominance, whereas lower scores indicate image dominance. Concepts with balanced scores act as shared cross-modality features.

Bridge Score. The bridge score $\mathbf{B} \in \mathbb{R}^{c \times c}$ quantifies the extent to which concepts serve as connectors between the image and the text modalities. It is defined as:

$$\mathbf{B} = \mathbb{E}_{(z_\iota, z_\tau) \sim \gamma} [z_\iota^{\top} z_\tau] \odot (DD^{\top}), \tag{5}$$

where $(z_\iota, z_\tau) \sim \gamma$ denotes the pair of sparse codes obtained from a matching image-text pair, and $\odot$ denotes the Hadamard product. Higher bridge scores indicate concepts that function as semantic connectors across the two modalities.

Retrieval Attribution Score. The retrieval attribution score $\mathbf{A} \in \mathbb{R}^{c}$ measures the contribution of concepts to the overall cross-modality similarity between image and text representations. The contribution of each concept is decomposed through reconstruction interactions:

$$\mathbf{A} = \mathbb{E}_{(z_\iota, z_\tau) \sim \gamma} \left[z_\iota \odot (M z_\tau) + z_\tau \odot (M z_\iota) \right], \tag{6}$$

where $M = DD^\top$. The score is tailored for multimodal retrieval settings and measures each concept's impact on cross-modality matching, with higher scores indicating greater influence on the final similarity between queries and candidates.

3.2 Observation: Textual information dominates representations in MLLMs

Key Takeaway

The strong text bias in MLLMs' representational space limits the encoding of visual information and undermines multimodal retrieval effectiveness.

We first analyze the distribution of the ***modality score*** (Eq. 4). In all models that we analyze, as shown in Figure 2, our first observation is that the majority of the learned concepts are single-modality. Furthermore, the concepts learned by MLLMs exhibit a strong bias towards the text modality. The distribution of modality scores reveals that a large proportion of concepts are text-specific, indicating that the representations generated by MLLMs are primarily driven by linguistic information rather than visual information.

However, unlike MLLMs, the distribution of image- and text-specific concepts is more balanced in CLIP and SigLIP2. Moreover, they exhibit a larger fraction of intermediate concepts, which encode information from both modalities, serving as shared semantic anchors that better connect the visual and textual representations.

Previous studies have discussed the existence of modality bias (Zheng et al., 2025) and modality gap (Liang et al., 2022; Eslami and de Melo, 2025) in multimodal models. In the context of multimodal retrieval, we argue that maintaining a balanced modality representation is particularly important. When a model's representations are heavily biased towards one modality, its embeddings fail to adequately represent samples from the other, leading to degraded retrieval performance under cross-modality or mixed-modality retrieval scenarios. Meanwhile, contrastive VLMs, with more balanced and interconnected concept spaces, are more capable of producing modality-agnostic embeddings, achieving better performance in zero-shot multimodal retrieval settings than MLLMs.

3.3 Observation: MLLMs concentrate most of their representational efforts on bridging image-text modalities

Key Takeaway

MLLMs overly align visual information into the text space, producing embeddings that are less distinctive across samples and impairing multimodal retrieval performance.

When we examine the ***energy*** (Eq. 3) distribution of the learned concepts, it is observed that for all models, the distribution exhibits a pronounced long-tail pattern (Figure 3). This reveals that during the reconstruction process, only a small subset of concepts is repeatedly activated across different samples. These concepts constitute the main concentrations of the model's representational space. These concepts therefore play a critical role in shaping the overall embedding structure.

To further explore where this representational energy is allocated, we analyze two additional metrics, ***bridge score*** (Eq. 5) and ***retrieval attribution score*** (Eq. 6), for a combined assessment. Our goal is to uncover how multimodal models, particularly MLLMs, distribute their energy across concepts and how such allocation patterns affect their retrieval performance. Specifically, for each concept, we compute its energy, bridge score, and retrieval attribution score, and then extract the top 1% concepts for each metric. We measure the Jaccard similarity between these top sets to quantify the degree of overlap, where a higher overlap demonstrates that the same set of concepts dominates multiple representational dimensions.

Formally, for two concept sets $\mathcal{A}$ and $\mathcal{B}$, the Jaccard similarity is computed as:

$$J(\mathcal{A}, \mathcal{B}) = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|}. \tag{7}$$

As shown in Table 1, both MLLMs and contrastive VLMs exhibit significant overlaps between the concepts of high-energy, high-bridge, and high-retrieval-attribution score sets. However, the overlap ratios are consistently higher in MLLMs. Notably, the intersection between high-energy and high-bridge concepts is particularly strong. This suggests that MLLMs devote a large portion of their representational energy to bridging modalities, attempting to harmonize image and text representations within a unified latent space.

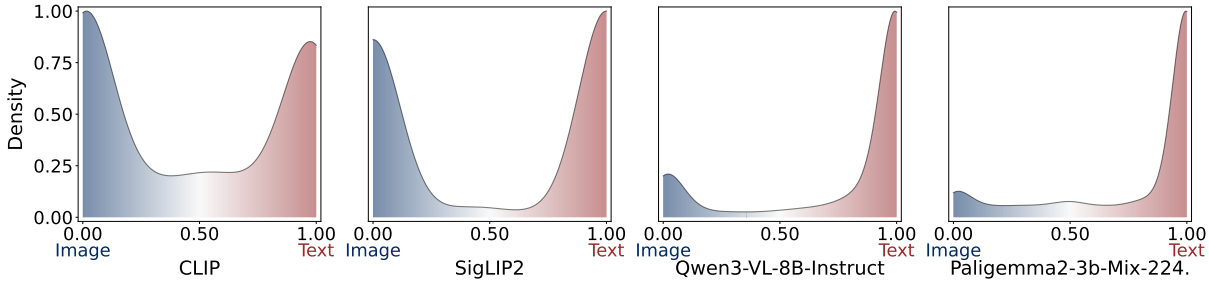

Figure 2: Distribution of ***modality scores*** for learned concepts by **(a) CLIP**, **(b) SigLIP2**, **(c) Qwen3-VL-8B-Instruct**, and **(d) Paligemma2-3B-Mix-224**. The Modality Score quantifies the bias of each concept towards the image modality (blue region) or the text modality (red region). The distributions are visualized using the Kernel Density Estimation (KDE) method (Parzen, 1962) based on concept activation statistics.

Type	**Model**	$J(\mathcal{S}_E, \mathcal{S}_A)$	$J(\mathcal{S}_E, \mathcal{S}_B)$	$J(\mathcal{S}_A, \mathcal{S}_B)$	**Triple Overlap**
Contrastive VLMs	CLIP	0.5674	0.6831	0.6126	0.4796
	SigLIP2	0.3948	0.5775	0.3783	0.2915
MLLMs	Qwen2-VL-7B-Instruct	0.6494	0.7957	0.7146	0.5989
	Qwen3-VL-8B-Instruct	0.6610	0.8547	0.6700	0.5486
	Paligemma2-3B-Mix-224	0.7655	0.8085	0.7124	0.6541
	Chameleon-7B	0.6850	0.8591	0.7099	0.5953

Table 1: Jaccard similarity between top 1% concept sets across three different evaluation metrics. $\mathcal{S}_E$: top ***energy*** set, $\mathcal{S}_A$: top ***retrieval attribution score*** set, $\mathcal{S}_B$: top ***bridge score*** set.

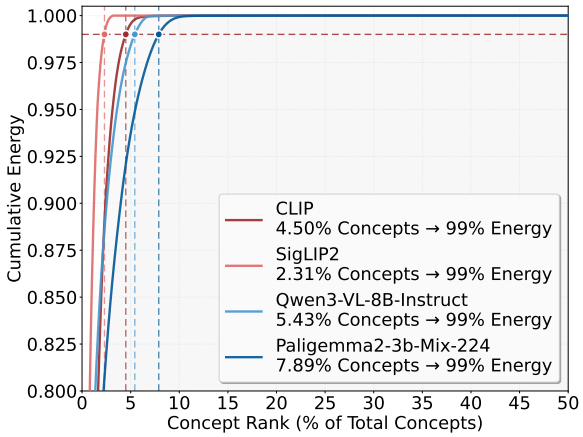

Figure 3: Cumulative ***energy*** distribution across concept ranks for different multimodal models. The curves show the percentage of total energy captured as concepts are ranked by their individual energy values.

At first glance, this behavior might appear beneficial for multimodal integration and even for retrieval. However, when combined with the findings discussed in Section 3.2, a clearer picture emerges: MLLMs primarily align multimodal information by projecting visual information towards the text space, rather than establishing a balanced semantic fusion. This could lead to an inferior ability of MLLMs to encode visual information.

To validate this, we conduct an additional experiment. Before mean pooling, we mask out the hidden states corresponding to image tokens and then recompute the final embeddings.

As shown in Figure 4, we find that removing image tokens results in only marginal changes in retrieval performance, while masking the user prompt region leads to a sharp performance decline.

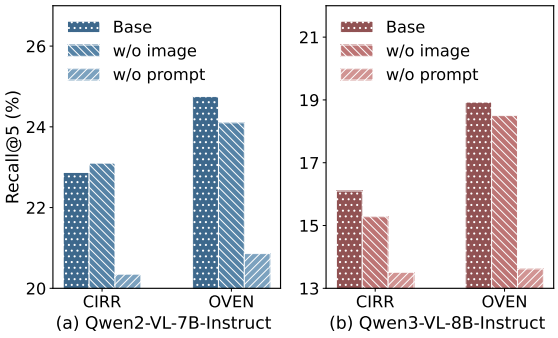

Figure 4: Retrieval performance on the subset (3k queries) of CIRR $((q_i, q_t) \rightarrow c_t)$ and OVEN $((q_i, q_t) \rightarrow (q_i, q_t))$. "Base" uses the full input; "w/o image" and "w/o prompt" denote the removal of image tokens and prompt tokens, respectively.

From a retrieval task perspective, effective retrieval performance largely depends on the distinctiveness of embeddings between samples. When we examine the overlap between high-retrieval-attribution score concepts (those most relevant to re-

trieval) and the high-energy / high-bridge score concepts, we observe a significant intersection among concepts learned from MLLMs. This suggests that the same dominant concepts not only consume most of the representational energy but also contribute most to multimodal similarity computation for retrieval. Consequently, the embeddings of different samples become concentrated around similar text-driven features with limited discriminability.

Meanwhile, contrastive VLMs display weaker overlap among these metrics. The high-energy, high-bridge score, and high-retrieval-attribution score concepts are more loosely coupled, allowing them to provide representations with greater discriminability. This ultimately contributes to their superior performance in multimodal retrieval.

3.4 Observation: Dominant components in similarity computation are counterproductive for retrieval

Key Takeaway

When MLLMs act as retrievers, the components exerting the greatest influence on similarity scores are in fact counterproductive distractors for multimodal retrieval.

Model	Setting	R@1	R@5	R@10
Qwen2-VL-7B	Base	5.60	15.44	22.66
	Removal	13.16	29.52	39.76
Qwen3-VL-8B	Base	4.71	13.86	21.34
	Removal	17.04	37.44	48.22
PaliGemma2-3B	Base	5.94	14.22	19.25
	Removal	23.18	46.84	58.64
Chameleon-7B	Base	1.79	7.58	10.12
	Removal	9.02	24.12	31.34

Table 2: Retrieval performance on the MSCOCO ($q_i \rightarrow c_t$) dataset before (Base) and after removing the subspace spanned by concepts with high retrieval attribution scores (Removal), evaluated on Qwen2-VL-7B (**Qwen2-VL-7B-Instruct**), Qwen3-VL-8B (**Qwen3-VL-8B-Instruct**), PaliGemma2-3B (**Paligemma2-3B-Mix-224**), and **Chameleon-7B**.

A core objective of retrieval is to compute a reliable similarity metric between query and candidate embeddings. Given the suboptimal performance of MLLMs in multimodal retrieval and the analysis in the previous section, a critical question arises: *Do the representation components that exert the greatest influence on the similarity calculation genuinely contribute to retrieval performance?*

To investigate this, we propose an ablation strategy that removes the components corresponding to concepts that exhibit high ***retrieval attribution scores*** (Eq. 6) from the representational space.

Concretely, let $D_R \subseteq D$ denote the sub-dictionary formed by the concept atoms whose retrieval attribution scores fall in the top 1% (m concepts). To characterize the subspace $\mathcal{S}$ spanned by these concepts with high influence, we perform Singular Value Decomposition (SVD) on D_R:

$$D_R = U \Sigma V^\top, \tag{8}$$

where $V \in \mathbb{R}^{d \times m}$ contains the right singular vectors that form an orthonormal basis for dominant directions in the embedding space. We select the top r right dominant singular vectors $V_r \in \mathbb{R}^{d \times r}$ to construct the basis for the subspace $\mathcal{S}$.

Given the embedding $h \in \mathbb{R}^d$, we project it onto $\mathcal{S}$ to isolate the components associated with D_R:

$$\Pi_{\mathcal{S}}(h) = V_r V_r^\top h. \tag{9}$$

We then compute the representation of removing the effect of D_R by subtracting this projection:

$$\tilde{h} = h - \Pi_{\mathcal{S}}(h). \tag{10}$$

The resulting residual embedding $\tilde{h}$ is normalized and utilized directly for retrieval.

The retrieval performance using the transformed embeddings is reported in Table 2, and the results reveal that removing the concepts with the highest retrieval attribution scores yields a significant improvement in retrieval accuracy.

This suggests that the components dominating the similarity calculation in MLLM-derived embeddings are, in effect, counterproductive for retrieval tasks. While these concepts possess heavy influence on the dot product magnitude, they appear to act as "distractors" that inflate similarity scores regardless of semantic relevance, implying that in the raw representational spaces of MLLMs, the semantically discriminative features necessary for precise retrieval are severely overshadowed.

4 ReAlign: An efficient test-time solution for MLLM multimodal retrieval

4.1 Methodology

Despite the vast parametric knowledge embedded within MLLMs, our empirical observations reveal

that their native representational space is suboptimal for multimodal retrieval, reflected as an imbalance in the distribution of semantic components within the embedding space. Under this analysis, a natural progression is seeking a transformation of the MLLM's output representation space that rectifies the feature distribution to enhance the semantic distinctiveness of query and candidate embeddings.

To this end, we propose ReAlign, a training-free *Test-Time Adaptation (TTA)* framework. ReAlign operates as a lightweight post-processing module that leverages a whitening strategy to realign the geometry of the feature space on-the-fly. This approach mitigates representational bottlenecks without requiring computationally expensive fine-tuning or access to labeled data.

Given the mean-pooled representations $\bar{h}_i \in \mathbb{R}^d$ derived from the last layer of MLLMs (as described in Section 2), ReAlign maps them into an isotropic space via a Zero-phase Component Analysis (ZCA) whitening strategy (Bell and Sejnowski, 1997). To ensure numerical stability in high-dimensional settings, we employ a shrinkage estimator for the covariance matrix calculation:

$$\hat{\Sigma} = (1-\beta)\Sigma + \beta \frac{\mathrm{Tr}(\Sigma)}{d} I, \tag{11}$$

where Σ is the empirical covariance of the centered features, $\beta \in [0,1]$ is the shrinkage coefficient, and I denotes the identity matrix. This regularization balances the empirical covariance with a spherical prior. Here, let $\hat{\Sigma} = U\Lambda U^\top$ be the eigen-decomposition. The final whitened and normalized representation e_i is computed as:

$$e_i = \frac{(\bar{h}_i - \mu)^\top U(\Lambda + \epsilon I)^{-\frac{1}{2}} U^\top}{\|(\bar{h}_i - \mu)^\top U(\Lambda + \epsilon I)^{-\frac{1}{2}} U^\top\|_2}, \tag{12}$$

where μ is the mean vector, ϵ (e.g., 10^{-5}) is a small constant for numerical stability, and $\|\cdot\|_2$ denotes ℓ_2-norm normalization. This transformation effectively aligns the embeddings to a spherical distribution, enhancing the semantic distinctiveness.

Specifically, we adopt an asymmetric estimation strategy to accommodate the distinct data flows of candidates and queries. For the candidate set, we compute the global statistics directly over the entire set to capture its holistic geometry. In contrast, we introduce a support set (sampled from the training set) to calculate the statistics for queries since queries arrive in mini-batches where local estimation suffers from high variance.

[1]The average results of three tests are reported

4.2 Experiments

4.2.1 Experimental setup

Benchmark. We evaluate the effectiveness of our proposed ReAlign on the M-BEIR benchmark (Wei et al., 2024). The benchmark comprises eight multimodal retrieval tasks involving both text and image modalities. Based on the datasets and task configurations (the modality combinations of queries and candidates), the tasks are further categorized into 16 distinct retrieval types. Additional details of the M-BEIR benchmark are provided in Appendix C.

MLLMs. We conduct a comprehensive evaluation across four representative MLLM architectures: Qwen2-VL-7B-Instruct (Wang et al., 2024), Qwen3-VL-8B-Instruct (Bai et al., 2025), Paligemma2-3B-Mix-224 (Steiner et al., 2024), and Chameleon-7B (Team, 2024).

4.2.2 Experimental results

Main Results. We evaluate zero-shot multimodal retrieval on the test set of the M-BEIR benchmark. Table 3 demonstrates the effectiveness of our proposed ReAlign. ReAlign consistently yields significant performance gains across all MLLM architectures. The improvements are particularly profound on large-scale datasets containing massive candidate items. In these scenarios, vanilla MLLMs often fail to produce distinctive embeddings, resulting in inferior retrieval results. For instance, on the InfoSeek dataset under the setting $((q_i, c_t) \rightarrow c_i)$ (Task 6), the vanilla Qwen3-VL-8B-Instruct model achieves a negligible Recall@5 of 0.79%. However, after applying ReAlign, the performance increases to 34.66%, marking an absolute improvement of 33.87%. Similar substantial gains can be observed in all other tasks.

Additionally, we observe that retrieval performance is not strictly correlated with the model's scale or general capabilities, suggesting that a larger MLLM does not necessarily imply superior representations for retrieval. For example, the performance of the 7B model, Chameleon-7B, does not surpass that of the 3B model, Paligemma2-3B-Mix-224, which may be attributed to the fact that these models are optimized for dense retrieval.

Comparison with Stronger Baselines. To further evaluate the effectiveness of ReAlign, we compare our method (based on Qwen2-VL-7B-Instruct) against two state-of-the-art specialized embedding models: GME (Zhang et al., 2024) and

Task	Dataset	Qwen2-VL-7B		Qwen3-VL-8B		Paligemma2-3B		Chameleon-7B	
		Base	ReAlign	Base	ReAlign	Base	ReAlign	Base	ReAlign
1. $q_t \to c_i$	VisualNews	2.45	15.46 +13.01	2.16	14.89 +12.73	3.97	12.01 +8.04	1.68	7.50 +5.82
	MSCOCO	18.57	59.80 +41.23	16.39	57.63 +41.24	26.63	66.38 +39.75	9.74	48.50 +38.76
	Fashion200K	6.98	9.37 +2.39	7.25	8.99 +1.74	1.37	5.06 +3.69	1.87	4.89 +3.02
2. $q_t \to c_t$	WebQA	18.98	64.59 +45.61	17.46	61.67 +44.21	21.37	60.34 +38.97	15.42	49.38 +33.96
3. $q_t \to (c_i, c_t)$	EDIS	7.13	31.84 +24.71	6.73	28.96 +22.23	9.54	25.16 +15.62	5.92	14.71 +8.79
	WebQA	2.19	66.43 +64.24	2.05	68.47 +66.42	4.62	64.95 +60.33	2.02	59.19 +57.17
4. $q_i \to c_t$	VisualNews	1.21	15.53 +14.32	1.18	15.05 +13.87	2.45	10.88 +8.43	1.05	6.09 +5.04
	MSCOCO	15.44	72.62 +57.18	13.86	66.18 +52.32	14.22	71.16 +56.94	9.18	44.64 +35.46
	Fashion200K	1.37	9.70 +8.33	1.41	12.49 +11.08	0.18	3.93 +3.75	0.15	4.36 +4.21
5. $q_i \to c_i$	NIGHTS	25.80	28.39 +2.59	25.14	27.71 +2.57	24.67	30.66 +5.99	19.16	27.09 +7.93
6. $(q_i, q_t) \to c_t$	OVEN	0.40	34.26 +33.86	0.35	33.65 +33.3	0.15	29.77 +29.62	0.02	25.31 +25.29
	InfoSeek	0.84	32.87 +32.03	0.79	34.66 +33.87	0.29	29.23 +28.94	0.18	16.52 +16.34
7. $(q_i, q_t) \to c_i$	FashionIQ	1.32	6.35 +5.32	1.03	6.74 +5.71	1.97	5.96 +3.99	0.89	4.84 +3.95
	CIRR	6.09	16.53 +10.44	4.46	15.94 +11.48	9.06	15.48 +6.42	7.19	10.36 +3.17
8. $(q_i, q_t) \to (c_i, c_t)$	OVEN	0.16	48.03 +47.87	0.14	41.65 +41.51	0.93	40.14 +39.21	0.25	34.57 +34.32
	InfoSeek	0.14	43.19 +43.05	0.09	41.11 +41.02	0.22	38.19 +37.97	0.08	22.94 +22.86
-	Average	6.82	34.69 +27.87	6.28	33.49 +27.21	7.60	31.83 +24.23	4.67	23.81 +19.14

Table 3: Multi-task evaluation on M-BEIR benchmark. We report Recall@10 for Fashion200K and FashionIQ, and Recall@5 for others[1]. "Base" refers to vanilla MLLMs, and "ReAlign" refers to the inclusion of our method (performance improvements are highlighted in red), evaluated across Qwen2-VL-7B (**Qwen2-VL-7B-Instruct**), Qwen3-VL-8B (**Qwen3-VL-8B-Instruct**), PaliGemma2-3B (**Paligemma2-3B-Mix-224**), and **Chameleon-7B**.

VLM2Vec (Jiang et al., 2024b). As shown in Table 4, these models are evaluated on four representative tasks from the M-BEIR benchmark.

Despite being training-free, ReAlign achieves competitive or even superior performance compared to these fine-tuned baselines. Notably, ReAlign outperforms both GME and VLM2Vec by a significant margin in WebQA. For instance, ReAlign reaches 66.43% in Recall@5 on WebQA, surpassing VLM2Vec and GME by 2.53% and 7.25% in absolute terms, respectively.

Task	Dataset	GME	VLM2Vec	ReAlign
1	MSCOCO	52.40	45.41	**59.80**
3	WebQA	59.18	63.90	**66.43**
4	MSCOCO	**86.70**	39.98	72.62
7	OVEN	39.11	**52.23**	48.03

Table 4: Comparison of ReAlign with fine-tuned embedding baselines on four tasks from M-BEIR. All models are based on Qwen2-VL-7B architecture. We report Recall@5, and **bold** values denote the best performance.

These results collectively validate that the representational spaces of MLLMs are a primary bottleneck for retrieval, and our proposed whitening transformation effectively aligns these representations to unlock the models' retrieval capabilities.

5 Related works

5.1 Multimodal large language models

Multimodal large language models (MLLMs) have achieved remarkable success in a wide range of tasks (Caffagni et al., 2024; Zhou et al., 2024). Prominent MLLMs such as LLaVA (Liu et al., 2023a, 2024a), Qwen-VL (Team, 2025), Paligemma2 (Steiner et al., 2024), InternVL (Wang et al., 2025), and MiniCPM-V (Yao et al., 2024) have shown promising advances in generative tasks such as image captioning, visual question answering, and complex multimodal reasoning (Sarto et al., 2025; Liang et al., 2025; Guo et al., 2025; Liu et al., 2024b; Liang et al., 2024).

5.2 Multimodal retrieval

Multimodal retrieval aims to align diverse modalities (e.g., text, images, or mixed text–image) within a shared embedding space for semantic similarity matching. Early multimodal retrieval approaches largely utilize pre-trained models such as CLIP (Radford et al., 2021) or BLIP (Li et al., 2022) for multimodal embedding. For instance, UniVL-DR (Liu et al., 2023b) and UniR (Wei et al., 2024) encode images and texts using CLIP or BLIP encoders separately, followed by fusion strategies.

With the advancement of MLLMs, researchers have begun to explore the potential of leveraging MLLMs in multimodal retrieval (Liu et al., 2025; Lan et al., 2025; Kong et al., 2025; Zhou et al., 2025). Specifically, a line of work focuses on building embedding models based on pretrained MLLMs to enhance multimodal retrieval, such as E5-V (Jiang et al., 2024a), VLM2Vec (Jiang et al., 2024b), MM-Embed (Lin et al., 2025), and CAFe (Yu et al., 2025). However, these methods often require substantial computational resources and typically rely on multi-stage training strategies (Zhang et al., 2025b; Huang et al., 2025), introducing considerable costs. In this paper, we are the first to systematically analyze the representational spaces of MLLMs from a multimodal retrieval perspective. Based on our observations, we propose a training-free approach that leverages intrinsic properties of MLLM embeddings to improve retrieval performance without requiring additional training.

5.3 Whitening transformations for representations

Whitening is a well-established technique for mitigating anisotropy in embedding spaces by decorrelating features and normalizing variance (Su et al., 2021; Diera et al., 2024; Liang et al., 2021). In natural language processing, methods such as BERT-flow (Li et al., 2020) and Whitening-BERT (Huang et al., 2021) apply whitening to transform BERT embeddings into a more isotropic distribution to improve semantic and representational properties. In this work, we extend the application of whitening to the more intricate high-dimensional representations of MLLMs.

6 Conclusion

In this work, we investigate the mechanisms that hinder MLLMs from functioning as effective zero-shot retrievers, despite their generative dominance. By leveraging sparse autoencoders (SAEs) to dissect the model's internal representations, we reveal that MLLMs suffer from a representational space overwhelmingly dominated by text, which suppresses essential visual information. Furthermore, we find that the heavy focus on bridging modalities for generation homogenizes embeddings, reducing the discriminative power required for retrieval, while the features most influential to similarity scores paradoxically act as distractors. Based on these insights, we propose ReAlign, a training-free test-time adaptation method that adjusts MLLM embedding geometry. This simple intervention consistently and substantially improves multimodal retrieval performance across different MLLMs.

7 Limitations

We acknowledge some limitations in our work. Firstly, due to limited computational resources, we are unable to experiment with larger open-source MLLMs, such as models with 30B or 70B parameters. Investigating how model scaling impacts the identified representational biases and the efficacy of our retrieval strategy would be a valuable direction for future research. Secondly, our investigation focuses primarily on the dominant image-text modality; extending the analysis and ReAlign strategy to other modalities, such as video or audio, remains a promising direction for future research.

Acknowledgments

This work is supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM113), National Natural Science Foundation of China (92470121, 62402016), National Key R&D Program of China (2024YFA1014003), Zhongguancun Academy (C20250204, C20250602), Beijing Major Science and Technology Project (Z251100008125043, Z251100008425023), and High-performance Computing Platform of Peking University.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Tianyi Bai, Hao Liang, Binwang Wan, Yanran Xu, Xi Li, Shiyu Li, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, and 1 others. 2024. A survey of multimodal large language model from a data-centric perspective. *arXiv preprint arXiv:2405.16640*.

Anthony J. Bell and Terrence J. Sejnowski. 1997. The "independent components" of natural scenes are edge filters. *Vision Research*, 37:3327–3338.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, and 1

others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *Annual Meeting of the Association for Computational Linguistics*.

Qifeng Cai, Hao Liang, Hejun Dong, Meiyi Qiang, Ruichuan An, Zhaoyang Han, Zhengzhou Zhu, Bin Cui, and Wentao Zhang. 2025. Lovr: A benchmark for long video retrieval in multimodal contexts. *arXiv preprint arXiv:2505.13928*.

Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. 2024. Understanding the interplay between parametric and contextual knowledge for large language models. *Preprint*, arXiv:2410.08414.

Andor Diera, Lukas Galke, and Ansgar Scherp. 2024. Isotropy matters: Soft-zca whitening of embeddings for semantic code search. *arXiv preprint arXiv:2411.17538*.

Sedigheh Eslami and Gerard de Melo. 2025. Mitigate the gap: Improving cross-modal alignment in CLIP. In *The Thirteenth International Conference on Learning Representations*.

Yan Fang, Jingtao Zhan, Qingyao Ai, Jiaxin Mao, Weihang Su, Jia Chen, and Yiqun Liu. 2024. Scaling laws for dense retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1339–1349, New York, NY, USA. Association for Computing Machinery.

Leo Gao, Tom Dupr'e la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *ArXiv*, abs/2406.04093.

Tianyu Guo, Hongyu Chen, Hao Liang, Meiyi Qiang, Bohan Zeng, Linzhuang Sun, Bin Cui, and Wentao Zhang. 2025. Brace: A benchmark for robust audio caption quality evaluation. *arXiv preprint arXiv:2512.10403*.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. Whiteningbert: An easy unsupervised sentence embedding approach. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 238–244.

Lang Huang, Qiyu Wu, Zhongtao Miao, and Toshihiko Yamasaki. 2025. Joint fusion and encoding: Advancing multimodal retrieval from the ground up. *CoRR*, abs/2502.20008.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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

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

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, and Neel Nanda. 2024. Interpreting attention layer outputs with sparse autoencoders. *arXiv preprint arXiv:2406.17759*.

Fanheng Kong, Jingyuan Zhang, Yahui Liu, Hongzhi Zhang, Shi Feng, Xiaocui Yang, Daling Wang, Yu Tian, Victoria W., Fuzheng Zhang, and Guorui Zhou. 2025. Modality curation: Building universal embeddings for advanced multimodal information retrieval. *arXiv preprint arXiv:2505.19650*.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. 2025. LLaVE: Large language and vision embedding models with hardness-weighted contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13721–13735, Suzhou, China. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9119–9130.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

Hao Liang, Zirong Chen, and Wentao Zhang. 2024. Evqascore: Efficient video question answering data evaluation. *arXiv preprint arXiv:2411.06908*.

Hao Liang, Meiyi Qiang, Yuying Li, Zefeng He, Yongzhen Guo, Zhengzhou Zhu, Wentao Zhang, and Bin Cui. 2025. Mathclean: A benchmark for synthetic mathematical data cleaning. *arXiv preprint arXiv:2502.19058*.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: understanding the modality gap in multi-modal contrastive representation learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Yuxin Liang, Rui Cao, Jie Zheng, Jie Ren, and Ling Gao. 2021. Learning to remove: Towards isotropic pre-trained bert embedding. In *International conference on artificial neural networks*, pages 448–459. Springer.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. Mm-embed: Universal multimodal retrieval with multimodal LLMS. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. 2025. Lamra: Large multimodal model as your advanced retrieval assistant. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 4015–4025. Computer Vision Foundation / IEEE.

Zheng Liu, Hao Liang, Xijie Huang, Wentao Xiong, Qinhan Yu, Linzhuang Sun, Chong Chen, Conghui He, Bin Cui, and Wentao Zhang. 2024b. Synthvlm: High-efficiency and high-quality synthetic data for vision language models. *arXiv preprint arXiv:2407.20756*.

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2023b. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In *Proceedings of ICLR*.

Aleksandar Makelov, George Lange, and Neel Nanda. 2024. Towards principled evaluations of sparse autoencoders for interpretability and control. *arXiv preprint arXiv:2405.08366*.

Alireza Makhzani and Brendan Frey. 2014. k-sparse autoencoders. *Preprint*, arXiv:1312.5663.

Neel Nanda. 2023. Open Source Replication & Commentary on Anthropic's Dictionary Learning Paper.

Bruno A. Olshausen and David J. Field. 1997. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37:3311–3325.

Isabel Papadimitriou, Huangyuan Su, Thomas Fel, Naomi Saphra, Sham M. Kakade, and Stephanie Gil. 2025. Interpreting the linear structure of vision-language model embedding spaces. *CoRR*, abs/2504.11695.

Emanuel Parzen. 1962. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. 2024. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*.

Sara Sarto, Marcella Cornia, and Rita Cucchiara. 2025. Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives. *Preprint*, arXiv:2503.14604.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. 2024. Paligemma 2: A family of versatile vlms for transfer. *Preprint*, arXiv:2412.03555.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXVII*, page 387–404, Berlin, Heidelberg. Springer-Verlag.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Qingyu Yin, Chak Tou Leong, Minjun Zhu, Hanqi Yan, Qiang Zhang, Yulan He, Wenjie Li, Jun Wang, Yue Zhang, and Linyi Yang. 2025. Constrain alignment with sparse autoencoders. *Preprint*, arXiv:2411.07618.

Hao Yu, Zhuokai Zhao, Shen Yan, Lukasz Korycki, Jianyu Wang, Baosheng He, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, and Hanchao Yu. 2025. Cafe: Unifying representation and generation with contrastive-autoregressive finetuning. *ArXiv*, abs/2503.19900.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2025a. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025b. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9274–9285.

Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, Linfeng Zhang, Danda Pani Paudel, Xuanjing Huang, Yu-Gang Jiang, Nicu Sebe, Dacheng Tao, Luc Van Gool, and Xuming Hu. 2025. Mllms are deeply affected by modality bias. *Preprint*, arXiv:2505.18657.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. 2025. 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)*, pages 19076–19095, Vienna, Austria. Association for Computational Linguistics.

Minxuan Zhou, Hao Liang, Tianpeng Li, Zhiyu Wu, Mingan Lin, Linzhuang Sun, Yaqi Zhou, Yan Zhang, Xiaoqin Huang, Yicong Chen, and 1 others. 2024. Mathscape: Evaluating mllms in multimodal math scenarios through a hierarchical benchmark. *arXiv preprint arXiv:2408.07543*.

A Comparison of embedding extraction strategy for MLLMs

In this section, we explore strategies for extracting embeddings using MLLMs. We primarily focus on four strategies: (1) **Last token**: This method is widely adopted in the existing literature (Liu et al., 2025). In this scenario, the multimodal input is typically constructed as *"<image> <text> Summarize the above image and sentence in one word: <emb>"*, where *"<image>"* and *"<text>"* serve as placeholders. The hidden representation of the final token, denoted as *"<emb>"*, is extracted as the embedding. While this strategy has proven effective when MLLMs are optimized for representation learning objectives, it may be suboptimal for zero-shot multimodal retrieval. (2) **Max pooling**: This strategy applies max pooling to the representations from the last layer's hidden states, selecting the maximum value across the sequence dimension. (3) **Min pooling**: This method constructs the embedding by selecting the minimum value across

Model	Setting	MSCOCO			NIGHTS			CIRR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Qwen2-VL-7B-Instruct	Last token	0.20	2.65	3.81	1.89	3.64	5.88	0.54	1.49	3.23
	Max pooling	1.84	6.24	10.16	4.43	17.68	28.49	0.84	4.45	7.68
	Min pooling	1.46	4.76	7.30	4.38	16.51	25.28	0.73	3.95	6.84
	Mean pooling	5.90	15.44	22.66	7.55	22.80	39.11	1.41	6.09	9.47
Qwen3-VL-8B-Instruct	Last token	0.34	2.49	3.86	1.05	2.96	5.75	0.49	1.35	3.38
	Max pooling	1.45	5.38	9.15	3.24	14.38	27.65	0.82	2.13	5.94
	Min pooling	1.21	4.60	6.46	2.83	12.40	25.84	0.66	1.98	5.01
	Mean pooling	4.71	13.86	21.34	6.46	20.14	36.42	1.01	4.46	7.03
Paligemma2-3B-Mix-224	Last token	1.63	3.23	5.81	1.29	4.26	8.42	0.43	1.84	4.16
	Max pooling	3.26	7.54	12.37	5.13	10.84	17.41	0.81	4.74	9.50
	Min pooling	2.75	5.18	9.84	3.74	7.10	13.76	0.64	3.17	8.02
	Mean pooling	5.94	14.22	19.25	8.96	24.67	44.79	1.13	9.06	13.74

Table 5: Zero-shot multimodal retrieval performance comparison between four embedding extraction strategies: **last token**, **max pooling**, **min pooling** and **mean pooling** on the MSCOCO ($(q_i \rightarrow c_t)$), NIGHTS ($q_i \rightarrow c_t$), and CIRR ($(q_i, q_t) \rightarrow c_i$) datasets.

the sequence dimension of the last layer's hidden states. (4) **Mean pooling**: This strategy applies mean pooling to the representations derived from the last layer's hidden states to obtain a comprehensive embedding for the entire sequence.

To evaluate these methods, we conduct comparative experiments on zero-shot multimodal retrieval.

As shown in Table 5, the mean pooling strategy consistently outperforms all other methods across datasets and architectures. This stands in sharp contrast to the last token strategy, which yields significantly inferior results. While max and min pooling improve upon the last token baseline by aggregating information across the sequence, they still fall short of mean pooling. Based on these findings, we adopt mean pooling over the last layer's hidden states as the default embedding extraction method for all MLLM-based retrieval experiments in this work, unless otherwise specified.

B Training details for SAEs

In this section, we provide the comprehensive training configuration used for the sparse autoencoders (SAEs) discussed in this paper.

In experiments, we employ Top-K SAEs (Gao et al., 2024) to learn interpretable concept representations from the activations of the COCO (Lin et al., 2014) dataset [2]. The training process involves processing billions of activations. For instance, the training run of Qwen2-VL-7B-Instruct encompasses approximately 28 billion activations.

For MLLMs, we train SAEs on Qwen2-VL-7B-Instruct (Wang et al., 2024), Qwen3-VL-8B-Instruct (Bai et al., 2025), Paligemma2-3B-Mix-224 (Steiner et al., 2024) and Chameleon-7B (Team, 2024). For these architectures, SAEs are trained on the hidden states of the last layer. These SAEs differ in input dimension, but share a fixed dictionary width of 32768. For contrastive VLMs, we utilize CLIP (ViT-Base-Patch32) (Radford et al., 2021) and SigLIP2(SigLip2-Base-Patch16-512) (Tschannen et al., 2025). The SAEs for these models are trained on the derived embeddings with a dictionary width of 7168.

We train SAEs using the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rate is set to $8e-4$ for all SAEs with a batch size of 4096. Our implementation utilizes the `Overcomplete`[3] framework. We conduct the training process on four NVIDIA H20 GPUs.

To mitigate the prohibitive RAM and VRAM costs associated with storing pre-computed embeddings for the entire dataset, we adopt a dynamic encoding strategy. Instead of pre-calculating all activations, the models encode a small proportion of the dataset on-the-fly, loading batches sequentially into GPU memory to derive the necessary activations. To ensure training stability and prevent the optimization process from learning temporal correlations present in sequential data, we maintain a shuffled buffer of these activations, following (Nanda, 2023). The SAEs sample training batches from this randomized buffer rather than

[2] https://huggingface.co/datasets/lmms-lab/COCO-Caption

[3] https://github.com/KempnerInstitute/overcomplete

Task	Dataset	Instruction (shown 1 out of 4)	Domain	Train	Dev	Test	Pool
1. $q^t \rightarrow c^i$	VisualNews	Identify news-related image match with the description	News	99K	20K	20K	542K
	MSCOCO	Find an everyday image match with caption	Misc.	100K	24.8K	24.8K	5K
	Fashion200K	Based on fashion description, retrieve matched image	Fashion	15K	1.7K	1.7K	201K
2. $q^t \rightarrow c^t$	WebQA	Find an paragraph from Wikipedia to answer the question	Wiki	16K	1.7K	2.4K	544K
3. $q^t \rightarrow (c^i, c^t)$	EDIS	Find a news image matching with the caption	News	26K	3.2K	3.2K	1M
	WebQA	Find a Wiki image that answer the question	Wiki	17K	1.7K	2.5K	403K
4. $q^i \rightarrow c^t$	VisualNews	Provide a news-related caption for the displayed image	News	100K	20K	20K	537K
	MSCOCO	Find a caption describe the an image	Misc.	113K	5K	5K	25K
	Fashion200K	Find a description for the fashion item in the image	Fashion	15K	4.8K	4.8K	61K
5. $q^i \rightarrow c^i$	NIGHTS	Find an image that is identical to the given image	Misc.	16K	2K	2K	40K
6. $(q^i, q^t) \rightarrow c^t$	OVEN	Retrieve a Wiki text that answer the given query about the image	Wiki	150K	50K	50K	676K
	InfoSeek	Find an article that answers the given question about the image	Wiki	141K	11K	11K	611K
7. $(q^i, q^t) \rightarrow c^i$	FashionIQ	Find an image to match the fashion image and style note	Fashion	16K	2K	6K	74K
	CIRR	I'm looking for a similar everyday image with the described changes	Misc.	26K	2K	4K	21K
8. $(q^i, q^t) \rightarrow (c^i, c^t)$	OVEN	Find a Wiki image-text pair to answer a question regarding an image	Wiki	157K	14.7K	14.7K	335K
	InfoSeek	Find a Wiki image-text pair to answers my question about this image	Wiki	143K	17.6K	17.6K	481K
	10 datasets	64 instructions	4 domains	1.1M	182K	190K	5.6M

Table 6: Summary of the M-BEIR benchmarks.

directly from the sequential stream.

C Details about M-BEIR dataset

We present the details of the M-BEIR benchmark (Wei et al., 2024) in Table 6. The benchmark comprises eight multimodal retrieval tasks involving both text and image modalities. It is important to note that the M-BEIR benchmark applies additional processing to the datasets it incorporates, which may result in differences from the standard evaluation of individual datasets. In this paper, we only utilize the test set for our evaluation.

D Details of evaluating learned concepts

In Section 3, we employ sparse autoencoders (SAEs) to conduct a mechanistic analysis of MLLMs' performance in multimodal retrieval. While SAEs are trained on the hidden states at the token granularity to capture the fundamental feature dictionary of the model's representational space, the retrieval task itself is performed at the sample level. In our work, the representation of a sample is typically derived via mean pooling: $\bar{h} = \frac{1}{n}\sum_{t=1}^{n} h_t$, where $h_t \in \mathbb{R}^d$ denotes the embedding of the t-th token.

To ensure that our analysis is directly aligned with the multimodal retrieval task, we apply SAEs to these aggregated sample-level embeddings $\bar{h}$. Since the retrieval similarity score is calculated using the inner product of the pooled vectors in our experiments (see Appendix A), any faithful attribution of this score to latent concepts must decompose the pooled vectors themselves. Meanwhile, although the SAE activation function is nonlinear, the encoder's affine transformation is linear. Thus, the pre-activation state of a pooled embedding $\bar{h}W_{enc}^{\top}+b$ effectively represents the arithmetic mean of token-level embeddings. Applying SAE to individual tokens would fail to account for how features interact during the pooling process, which is critical to understanding retrieval failures.

E Efficiency analysis of ReAlign

In order to investigate the efficiency and scalability of ReAlign, we conduct an analysis on the EDIS $(q^t \rightarrow (c^i, c^t))$ dataset, which contains 1 million candidate items.

We compare the computational cost of our proposed module against the downstream similarity computation stage. It is important to note that a direct large-scale matrix multiplication for 1 million candidates is computationally prohibitive. Therefore, in our implementation, the similarity computation is optimized via batch processing. Specifically, we compute similarity matrices in small batches, extract the Top-K indices for each batch, and finally aggregate the results for evaluation.

As observed in Table 7, the computational overhead of ReAlign is comparable to that of similarity search. Even as the embedding dimension increases, our framework does not exhibit significant computational bottlenecks. The results demonstrate that ReAlign maintains acceptable time and memory costs, validating its efficiency and scala-

Model	Hidden Dim	ReAlign		Similarity Computation	
		Time (ms)	Mem (MiB)	Time (ms)	Mem (MiB)
PaliGemma2-3B	2,304	1,538	43,577	4,107	43,577
Qwen2-VL-Instruct-7B	3,584	3,623	58,978	5,162	44,677
Qwen3-VL-Instruct-7B	4,096	4,533	65,138	5,369	48,778

Table 7: Efficiency analysis on the EDIS dataset (containing 1M candidates). We report the computation time and peak memory usage across utilizing diverse MLLMs as retrievers. The results show that ReAlign introduces acceptable computation cost.

bility for large-scale retrieval tasks.

F Which MLLM layer contributes the most to multimodal retrieval?

In this section, we analyze the contribution of each layer of MLLM to multimodal retrieval performance. Specifically, we extract the embeddings from the hidden states of individual layers and evaluate their retrieval performance independently.

As shown in Figure 5, retrieval performance generally improves as representations are extracted from deeper layers, although progression varies between different architectures. For Qwen2-VL-7B-Instruct, we observe a distinct performance spike in the initial layers, indicating that semantic information is captured earlier. However, their performance remains limited compared to representations derived from deeper layers.

For both models (Qwen2-VL-7B-Instruct and Paligemma2-3B-Mix-224), embeddings from deeper layers consistently achieve higher recall across all evaluation metrics (R@1, R@5 and R@10). In particular, the last layer representations yield the best retrieval performance. Based on these observations, we adopt the hidden states from the final layer as the default representation for multimodal retrieval in all experiments.

G Parameter sensitivity analysis

We investigate how the performance of ReAlign varies with respect to the parameter β, which controls the shrinkage coefficient in the covariance estimation. The evaluation is conducted across the MSCOCO and CIRR datasets using two MLLM architectures under the zero-shot multimodal retrieval. The results, illustrated in Figure 6, map the retrieval performance (R@1, R@5 and R@10) as β varies from 0 to 1 in intervals of 0.1.

As shown in Figure 6, performance remains consistently strong within the moderate range (e.g., [0.1, 0.6]) but deteriorates at the extremes. This

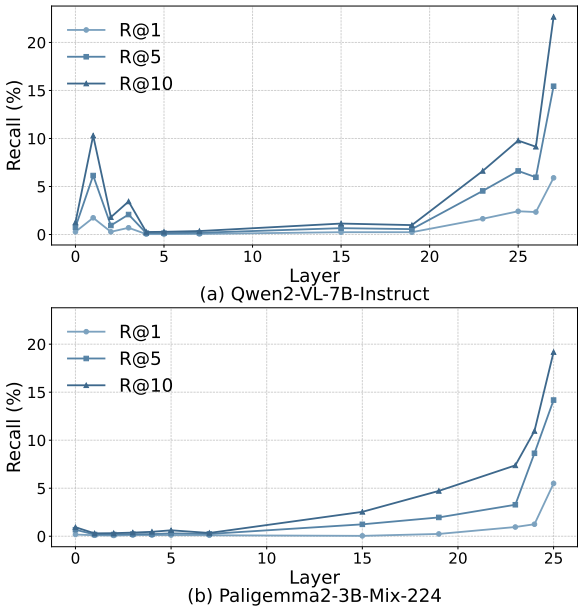

Figure 5: Retrieval performance of MLLMs across different layers on the MSCOCO ($q^{\text{i}} \rightarrow c^{\text{t}}$) dataset.

indicates that while a purely empirical covariance ($\beta = 0$) may suffer from estimation noise and a purely spherical prior ($\beta = 1$) discards critical semantic correlations, a balanced regularization effectively stabilizes the feature geometry.

H Comparison of ReAlign with standard PCA and ZCA in multimodal retrieval

In this section, we analyze the effectiveness of ReAlign by comparing it against the standard Principal Component Analysis (PCA) and Zero-phase Component Analysis (ZCA) baselines. As shown in Table 8, ReAlign consistently outperforms the standard PCA and ZCA across various MLLM backbones and datasets.

The performance gap stems primarily from the estimation strategy of covariance statistics Σ: The standard PCA and ZCA rely on the raw empirical covariance matrix Σ. If the embedding dimension is high (e.g., 4096), the whitening transformation can aggressively amplify noise along the directions

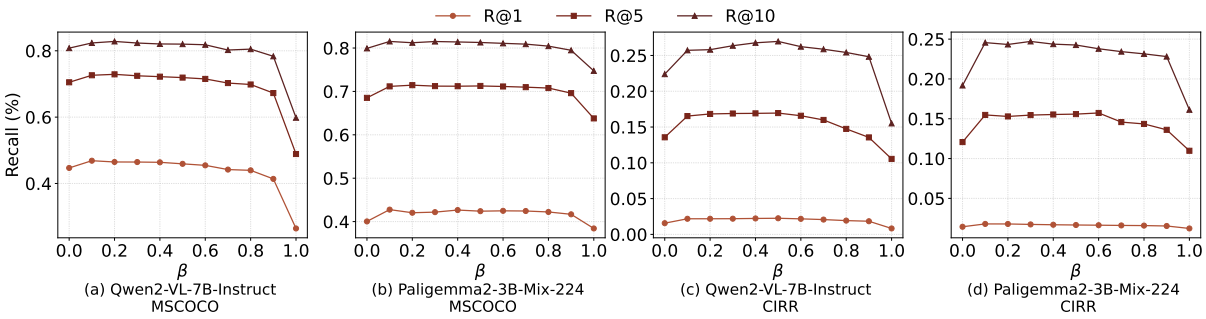

Figure 6: Impact of the shrinkage coefficient β on retrieval performance across the MSCOCO $(q_i \rightarrow c_t)$ and CIRR $((q_i, q_t) \rightarrow c_i)$ datasets.

corresponding to small eigenvalues, leading to unstable feature rectification. Our approach mitigates these issues by employing a shrinkage estimator, interpolating between the empirical covariance and a spherical prior (identity matrix). Empirical results demonstrate that ReAlign significantly improves retrieval performance.

I Comparison with contrastive VLMs

While our primary objective is to strengthen the multimodal retrieval capabilities of MLLMs without parameter updates, comparing ReAlign with specialized contrastive VLMs (e.g., CLIP and SigLIP) offers valuable insights into the distinct advantages of different architectural paradigms.

As shown in Table 9, contrastive models generally maintain a performance edge on standard image-to-text or text-to-image matching tasks (e.g., VisualNews, MSCOCO). This is expected, as these models are explicitly pre-trained with contrastive objectives (InfoNCE) designed to maximize the cosine similarity between text-image pairs. However, ReAlign significantly narrows this gap, transforming MLLMs from near-random retrievers into competitive baselines.

More importantly, our approach reveals the unique strength of MLLMs in complex and fused retrieval scenarios, especially in tasks requiring composite queries and candidates ($(q_t \rightarrow (c_i, c_t)), ((q_i, q_t) \rightarrow c_t)$) and $((q_i, q_t) \rightarrow (q_i, q_t))$. MLLMs equipped with ReAlign frequently outperform contrastive baselines (e.g., Qwen2-VL with ReAlign achieves 48.03 on OVEN Task 8 vs. 38.8 for CLIP). This suggests that ReAlign successfully leverages the deep semantic comprehension capabilities inherent in MLLMs, allowing them to handle complex retrieval tasks. Thus, ReAlign does not serve to replace specialized retrievers, but rather to equip generative giants with a robust and effective training-free retrieval mechanism.

[4]The average results of three tests are reported

Model	Setting	MSCOCO			NIGHTS			CIRR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Qwen2-VL-7B-Instruct	PCA	38.05	61.67	73.34	4.31	20.21	39.79	0.61	9.86	16.14
	ZCA	44.58	70.50	80.80	7.75	26.95	45.53	1.57	13.57	22.39
	ReAlign	46.86	72.62	82.36	7.83	28.40	47.69	2.18	16.53	25.71
Qwen3-VL-8B-Instruct	PCA	30.22	59.13	73.82	4.11	22.28	37.36	0.19	7.24	14.81
	ZCA	35.57	63.71	76.02	6.69	26.14	41.90	0.86	14.78	20.19
	ReAlign	38.96	66.18	77.36	7.83	27.71	45.83	1.51	15.94	24.56
Paligemma2-3B-Mix-224	PCA	37.43	62.11	77.30	5.75	27.46	49.83	2.01	19.79	45.13
	ZCA	40.71	69.22	79.15	9.74	27.74	48.45	6.62	26.78	43.16
	ReAlign	42.76	71.16	81.52	9.25	30.66	49.77	9.26	30.66	49.77
Chameleon-7B	PCA	12.92	33.54	50.47	5.16	22.78	39.64	0.15	8.07	13.94
	ZCA	19.41	41.56	56.02	6.98	24.86	41.03	0.09	7.76	14.11
	ReAlign	21.38	44.64	57.04	7.07	27.09	46.79	0.84	10.36	15.78

Table 8: Performance comparison between the standard PCA, ZCA baselines and our proposed ReAlign on the MSCOCO ($(q_i \to c_t)$), NIGHTS ($q_i \to c_t$), and CIRR ($(q_i, q_t) \to c_i$) datasets.

		Contrastive VLMs		Qwen2-VL-7B		Qwen3-VL-8B		Paligemma2-3B		Chameleon-7B	
Task	**Dataset**	CLIP	SigLIP	Base	ReAlign	Base	ReAlign	Base	ReAlign	Base	ReAlign
1. $q_t \to c_i$	VisualNews	43.3	30.1	2.45	15.46	2.16	14.89	3.97	12.01	1.68	7.50
	MSCOCO	61.1	75.7	18.57	59.80	16.39	57.63	26.63	66.38	9.74	48.5
	Fashion200K	6.6	36.5	6.98	9.37	7.25	8.99	1.37	5.06	1.87	4.89
2. $q_t \to c_t$	WebQA	36.2	39.8	18.98	64.59	17.46	61.67	21.37	60.34	15.42	49.38
3. $q_t \to (c_i, c_t)$	EDIS	43.3	27.0	7.13	31.84	6.73	28.96	9.54	25.16	5.92	14.71
	WebQA	45.1	43.5	2.19	66.43	2.05	68.47	4.62	64.95	2.02	59.19
4. $q_i \to c_t$	VisualNews	41.3	30.8	1.21	15.53	1.18	15.05	2.45	10.88	1.09	6.09
	MSCOCO	79.0	88.2	15.44	72.62	13.86	66.18	14.22	71.16	9.18	44.64
	Fashion200K	7.7	34.2	1.37	9.70	1.41	12.49	0.18	3.93	0.15	4.36
5. $q_i \to c_t$	NIGHTS	26.1	28.9	25.80	28.39	25.14	27.71	24.67	30.66	19.16	27.09
6. $(q_i, q_t) \to c_t$	OVEN	24.2	29.7	0.40	34.26	0.35	33.65	0.15	29.77	0.02	25.31
	InfoSeek	20.5	25.1	0.84	32.87	0.79	34.66	0.29	29.23	0.18	16.52
7. $(q_i, q_t) \to c_i$	FashionIQ	7.0	14.4	1.32	6.35	1.03	6.74	1.97	5.96	0.89	4.84
	CIRR	13.2	22.7	6.09	16.53	4.46	14.94	9.06	15.48	7.19	10.36
8. $(q_i, q_t) \to (c_i, c_t)$	OVEN	38.8	41.7	0.16	48.03	0.14	41.65	0.93	40.14	0.25	34.57
	InfoSeek	26.4	27.4	0.14	43.19	0.09	41.11	0.22	38.19	0.08	22.94
-	Average	32.5	37.2	6.82	34.69	6.28	33.49	7.60	31.83	4.67	23.81

Table 9: Multi-task evaluation on M-BEIR benchmark. We report Recall@5 for all datasets except Fashion200K and FashionIQ where Recall@10 is used[4].