Bridging the Modality Gap by Similarity Standardization with Pseudo-Positive Samples

Shuhei Yamashita, Daiki Shirafuji, Tatsuhiko Saito

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

Bridging the Modality Gap by Similarity Standardization with Pseudo-Positive Samples

Shuhei Yamashita **Daiki Shirafuji** **Tatsuhiko Saito**
Mitsubishi Electric Corporation
{Yamashita.Shuhei@bc, Shirafuji.Daiki@ay, Saito.Tatsuhiko@db}
.MitsubishiElectric.co.jp

Abstract

Advances in vision-language models (VLMs) have enabled effective cross-modality retrieval. However, when both text and images exist in the database, similarity scores would differ in scale by modality. This phenomenon, known as the modality gap, hinders accurate retrieval. Most existing studies address this issue with manually labeled data, e.g., by fine-tuning VLMs on them. In this work, we propose a similarity standardization approach with pseudo data construction. We first compute the mean and variance of the similarity scores between each query and its paired data in text or image modality. Using these modality-specific statistics, we standardize all similarity scores to compare on a common scale across modalities. These statistics are calculated from pseudo pairs, which are constructed by retrieving the text and image candidates with the highest cosine similarity to each query. We evaluate our method across seven VLMs using two multi-modal QA benchmarks (MMQA and WebQA), where each question requires retrieving either text or image data. Our experimental results show that our method significantly improves retrieval performance, achieving average Recall@20 gains of 64% on MMQA and 28% on WebQA when the query and the target data belong to different modalities. Compared to E5-V, which addresses the modality gap through image captioning, we confirm that our method more effectively bridges the modality gap.

1 Introduction

Information retrieval (IR) plays a key role in a wide range of NLP applications, including web search engines (Kobayashi and Takeda, 2000) and question answering systems (Kolomiyets and Moens, 2011). While traditional approaches primarily focus on retrieving textual information (Robertson and Zaragoza, 2009; Karpukhin et al., 2020), there is a growing interest in retrieving both text and images to provide richer and more informative results (Zhou et al., 2024b).

Vision-language models (VLMs), such as CLIP (Radford et al., 2021), enable both text and image data to be embedded into a shared representation space. Although VLMs enable effective text-to-image retrieval (Radford et al., 2021), it is still challenging to extract relevant information from a database that contains both text and images. Specifically, text items often dominate the top-ranked results even when relevant images exist (Chang et al., 2021; Liu et al., 2023). This issue is attributed to the *modality gap*—a phenomenon in which embeddings from different modalities are mapped to separate regions of the representation space (Liang et al., 2022). Consequently, data that share the same modality as the query tend to receive higher similarity scores, regardless of actual relevance (illustrated in Figure 1).

To address this problem, several approaches have been proposed. Some methods address the modality gap by fine-tuning pre-trained VLMs using paired datasets consisting of queries and their manually labeled corresponding text or image data (Fahim et al., 2024; Eslami and de Melo, 2025). Other methods for converting visual data into text have also been introduced, such as E5-V (Jiang et al., 2024). However, these approaches have shortcomings: collecting human-annotated data is resource-intensive, whereas image captioning would fail to preserve necessary visual information in text.

In this study, we propose a retrieval method that mitigates the impact of modality gap without manually labeled data or image captioning. The key idea is to make similarity scores comparable across modalities by standardizing them using the modality-specific mean and variance. To estimate these statistics, we construct pseudo-positive pairs of unlabeled queries and their most similar texts or images. We then derive modality-specific mean and variance from these pairs, which are used to

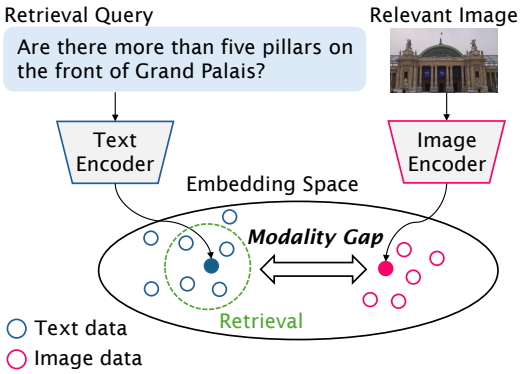


Figure 1: Conceptual overview of the modality gap. Texts and their corresponding images are projected to distant regions of the embedding space.

standardize similarity scores during retrieval.

To evaluate our approach, we conduct experiments on multi-modal question answering benchmarks, i.e., MMQA (Talmor et al., 2021) and WebQA (Chang et al., 2021) with seven pre-trained VLMs. Our method significantly improves retrieval performance when the query and the target data belong to different modalities, achieving average gains of 64% and 28% in Recall@20 on MMQA and WebQA, respectively.

Our main contributions are as follows:

- We propose a similarity standardization approach to mitigate the effect of the modality gap on multi-modal retrieval.
- Our method improves the retrieval performances on two datasets regardless of modalities, compared to E5-V.
- Our method bridges the modality gap without manually labeled datasets, such as pairs of queries and their corresponding examples.

2 Related Work

2.1 Multi-Modal Retrieval

Vision-language models (VLMs) have shown remarkable progress in recent years (Radford et al., 2021; Jia et al., 2021). These models are typically trained using contrastive learning to align images and text in a representation space. Their embeddings can be used for retrieval by computing similarity scores with each item in the database (Karpukhin et al., 2020).

Retrieval tasks involving multiple modalities can be broadly categorized into two settings (Liu et al., 2023). *Cross modality retrieval* refers to settings in which the query and target belong to different modalities, such as text-image or image-text retrieval. In contrast, *multi-modal retrieval* assumes that the retrieval database contains data from multiple modalities—for example, both text and images—and the goal is to find the most relevant item regardless of its modality.

While contrastively trained VLMs perform well in cross modality retrieval tasks (Radford et al., 2021), their performance in multi-modal retrieval remains limited. In particular, when both text and images are present in the retrieval set, these models often retrieve items only from the same modality as the query, and fail to retrieve relevant data from the other modality (Chang et al., 2021; Ross et al., 2024).

This issue is attributed to the modality gap, a clear separation between image and text embeddings of contrastively trained VLMs. This phenomenon was first studied by Liang et al. (2022), who showed that it exists even in randomly initialized models and persists throughout contrastive training. Several causes have been suggested in prior work, including an information imbalance between text and image inputs (Schrodi et al., 2025).

2.2 Bridging the Modality Gap

Some approaches attempt to eliminate the modality gap in VLMs by modifying the contrastive training process. (Fahim et al., 2024) augment CLIP's objective with uniformity and alignment regularizers to enforce balanced embedding distributions and eliminate the modality gap. Schrodi et al. (2025) demonstrated that contrastive learning can mitigate the modality gap when the training data is balanced in information content across modalities. Eslami and de Melo (2025) introduce AlignCLIP, which adds shared parameters between visual and text encoders and an intra-modality separation term to the contrastive loss. While effective, these methods require access to manually paired datasets, which can be expensive or unavailable in real-world scenarios.

Another line of work obtains image embeddings by leveraging image captions (Liu et al., 2023; Zhou et al., 2024a,b). These models achieve strong performance in multi-modal retrieval, but rely heavily on captions. In settings without image descriptions, retrieval quality deteriorates, indicating limited use of visual features.

More recently, methods utilizing the vision-language capabilities of multi-modal large language

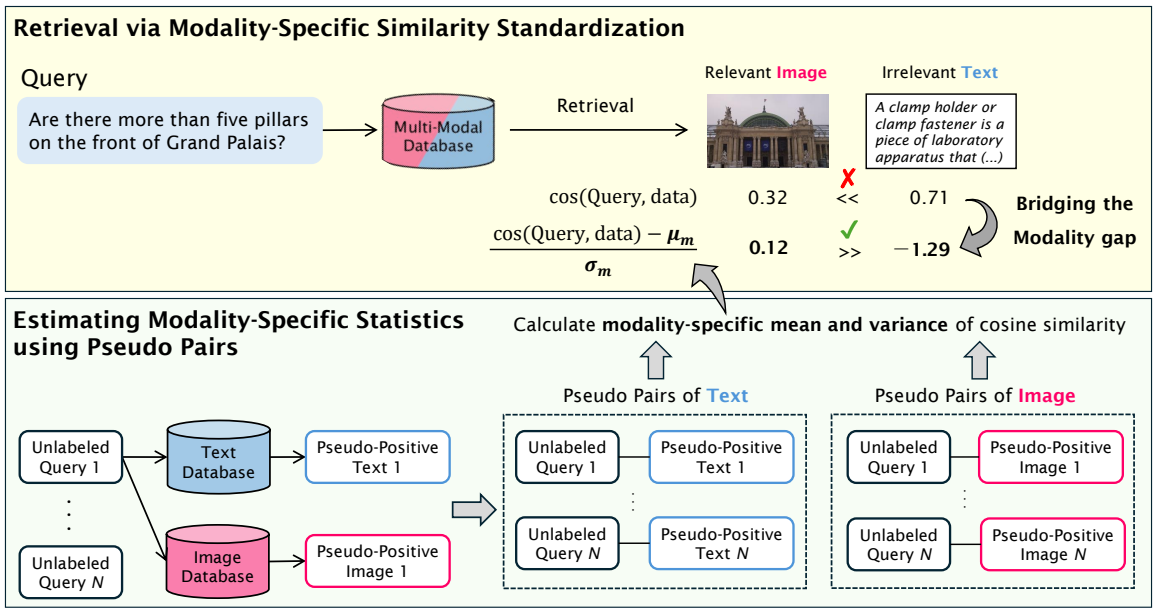


Figure 2: Overview of our proposed method. The modality gap causes irrelevant text to score higher than relevant images. Our approach addresses this issue by standardizing cosine similarity scores based on modality-specific mean and variance calculated from pseudo data.

models (MLLMs) have been explored (Jiang et al., 2024; Zhang et al., 2024b; Lin et al., 2025). For instance, E5-V (Jiang et al., 2024) prompts its backbone MLLM with an image to generate a one-word summary of it. By using the resulting features to obtain image embeddings, E5-V aligns visual inputs with the language space, effectively eliminating the modality gap.

Unlike the existing works that require manually labeled data or image captioning, our method directly adjusts similarity scores across modalities using pseudo-positive examples, eliminating the need for manual supervision.

3 Task Formulation

We work on the task of retrieving relevant data from a multi-modal database that contains both text and images, given a natural language query.

Formally, let q be a textual query and let $\mathcal{D} = \mathcal{D}_{\text{text}} \cup \mathcal{D}_{\text{image}}$ denote the retrieval database, where $\mathcal{D}_{\text{text}}$ and $\mathcal{D}_{\text{image}}$ are sets of textual and visual items respectively. A pre-trained VLM f encodes both the query and each item in the database into the same space. For each candidate $d \in \mathcal{D}$, its relevance to the query can be measured by comparing their embeddings, for example, using cosine similarity: $\cos(f(q), f(d))$.

However, due to the modality gap, similarities differ in scale between text and image modalities. Specifically, a text query tends to assign higher scores to textual candidates than to images, causing relevant images to appear lower in the ranking.

4 Proposed Methods

In this section, we propose a method that mitigates the negative impact of the modality gap without manually labeled data. We first introduce similarity standardization approach as described in Section 4.1. Then, we construct pseudo pairs instead of labeled data, as detailed in Section 4.2.

4.1 Modality-Specific Similarity Standardization

To bridge the modality gap, we propose a similarity standardization approach with modality-specific statistics. We standardize the similarity scores between queries and target information (i.e., positive examples) using their means and variances computed separately for text targets and image targets.

Let $\mathcal{P}_m$ be a set of query-positive pairs where positive example belongs to modality $m \in \{\text{text}, \text{image}\}$. We calculate the mean and variance

of similarities for each modality as:

$$\mu_m = \frac{1}{|\mathcal{P}_m|} \sum_{(q,d_m^+)\in\mathcal{P}_m} \cos(f(q), f(d_m^+)),$$
$$\sigma_m^2 = \frac{1}{|\mathcal{P}_m|} \sum_{(q,d_m^+)\in\mathcal{P}_m} (\cos(f(q), f(d_m^+)) - \mu_m)^2, \quad (1)$$

where each $(q, d_m^+) \in \mathcal{P}_m$ is a query-positive pair.

Using the modality-specific statistics estimated above, we standardize the cosine similarity between a query q and a candidate $d \in \mathcal{D}$ of modality m as:

$$\text{sim}(q, d) = \frac{\cos(f(q), f(d)) - \mu_m}{\sigma_m}. \quad (2)$$

This modality-aware standardization will mitigate the negative impact of the modality gap on similarities between text and image. Note that the statistics μ_m and σ_m^2 are computed from the pre-collected dataset $\mathcal{P}_m$ and remain fixed regardless of the retrieval queries.

4.2 Pseudo Pair Construction

We propose a method for constructing pseudo data that eliminates the need for manually labeled data.

Let $\mathcal{D}_m$ be the subset of the retrieval database corresponding to modality $m \in \{\text{text}, \text{image}\}$, and let $\mathcal{Q}$ denote a set of unlabeled queries. Given a query $q \in \mathcal{Q}$, we extract the most similar item from $\mathcal{D}_m$ for each modality m, and treat it as a pseudo-positive example of modality m:

$$\hat{d}_m^+ = \arg\max_{d\in\mathcal{D}_m} \cos(f(q), f(d)). \quad (3)$$

By repeating this process for all queries in $\mathcal{Q}$, we construct a modality-specific pseudo pair set $\hat{\mathcal{P}}_m$ for each modality m:

$$\hat{\mathcal{P}}_m = \{(q, \hat{d}_m^+) \mid q \in \mathcal{Q}\}. \quad (4)$$

$\hat{\mathcal{P}}_m$ can be used as a substitute for the manually labeled set $\mathcal{P}_m$ in Equations (1). This allows our method to perform modality-specific standardization without relying on any labeled data.

5 Experimental Setup

5.1 Datasets for Evaluation

We evaluate our method on two multi-modal question answering datasets: MultimodalQA (Talmor et al., 2021) and WebQA (Chang et al., 2021). These datasets are widely used benchmarks for the multi-modal retrieval task (Chen et al., 2022; Liu et al., 2023; Zhou et al., 2024a,b). In our experiments, we use questions that require retrieving relevant textual passages (TextQ) or images (ImageQ) in order to answer them. Table 1 shows examples from each dataset, and Table 2 shows the dataset sizes.

(a) MMQA: "*How many colors are on the Mississippi flag?*"

(b) WebQA: "*Are there more than five pillars on the front of Grand Palais?*"

Figure 3: Examples of positive images for ImageQ in MMQA and WebQA shown in Table 1.

MultiModalQA (MMQA) (Talmor et al., 2021) is a benchmark for multi-hop question answering across multiple modalities, including text, images, and tables. It is constructed from Wikipedia tables linked with relevant textual paragraphs and images via shared entities.

WebQA (Chang et al., 2021) is a large-scale open-domain question answering dataset that includes questions paired with corresponding textual passages or images. The data is collected from the open web and Wikipedia. Following Liu et al. (2023) and Zhou et al. (2024b), we construct a retrieval corpus by collecting all images and text passages relevant to all queries in the WebQA dataset.

5.2 Datasets for Pseudo Pair Construction

Pseudo pairs are constructed independently for the MMQA and WebQA datasets. We use queries from the training split of each dataset and sample their pseudo-positive examples from the retrieval source of each dataset as illustrated in Equation 3.

5.3 Metrics

We evaluate our methods using Recall@k, MRR@k, and NDCG@k. All metrics are primarily measured at $k = 20$. For Recall, we additionally compute values at k=1, 5, and 100 to examine the effect of varying k.

5.4 Models

We apply our method to seven pre-trained VLMs to demonstrate its robust effectiveness. To assess models expected to exhibit a modality gap due to contrastive training, we include CLIP (Rad-

Dataset	Type	Question	Positive Example
MMQA	TextQ	When did "Harry Potter and the Sorcerer's Stone" movie come out?	Harry Potter and the Philosopher's Stone (released in the United States as Harry Potter and the Sorcerer's Stone) is a 2001 fantasy film directed by Chris Columbus and distributed by Warner Bros.
	ImageQ	How many colors are on the Mississippi flag?	Refer to Figure 3a.
WebQA	TextQ	What part of the human body does the nerves in the frontalis muscle serve and the occipitofrontalis muscle serve?	The frontalis muscle is supplied by the facial nerve and receives blood from the supraorbital and supratrochlear arteries. In humans, the occipitofrontalis only serves for facial expressions.
	ImageQ	Are there more than five pillars on the front of Grand Palais?	Refer to Figure 3b.

Table 1: Examples from MMQA and WebQA datasets. Each dataset includes two types of questions: TextQ and ImageQ, which refer to questions that require retrieving text and images to answer, respectively.

# of dataset	Source		Query	
	text	image	TextQ	ImageQ
MMQA	218K	57K	6.7K/721	1.9K/230
WebQA	787K	389K	15K/2.4K	16K/2.5K

Table 2: Numbers of retrieval candidates and queries in MMQA and WebQA. The numbers of queries are listed as training/test. Validation data is not used in our experiments.

ford et al., 2021) (ViT-B/32 and ViT-L/14), LongCLIP (Zhang et al., 2024a) (base and large), and BLIP (Li et al., 2022). We also include Cohere Embed 3 English (Ross et al., 2024), a high-performance VLM accessible via API. In addition, we evaluate E5-V (Jiang et al., 2024), which integrates image captioning via a MLLM. While E5-V is designed to mitigate the modality gap, we apply similarity standardization to examine whether our method can further improve its performance. The computational resources are provided in Appendix A.

5.5 Evaluation Conditions

All VLMs are evaluated under the following three configurations.
(i) **Cos**: Cosine similarities are simply used for retrieval.
(ii) **Std**: Cosine similarities are standardized by our method with manually labeled data, which is taken from the training split of each dataset.
(iii) **Ours**: Cosine similarities are standardized by our method with our pseudo pairs.

6 Results and Discussions

6.1 Overall Results

Table 3 summarizes the overall retrieval performance across seven VLMs on MMQA and WebQA datasets. When Cos was applied, four of the CLIP-based models and BLIP retrieved almost no relevant results, resulting in near-zero scores on all evaluation metrics on ImageQ. This suggests that the modality gap causes irrelevant text passages to be ranked higher than relevant images, hindering accurate retrieval.

In contrast, applying our method to these models significantly improved the performances, achieving average gains of 64% and 28% in Recall@20 for MMQA ImageQ and WebQA ImageQ, respectively, thereby confirming its effectiveness in bridging the modality gap.

Notably, all models with our method outperformed E5-V on ImageQ. These results highlight the advantage of processing images without any loss of information, different from the existing works with image captioning or verbalization. Although a slight performance degradation was observed on TextQ, the overall trade-off is favorable with notable gains on ImageQ.

Cohere Embed 3 and E5-V achieved high performance on TextQ, with approximately 80% in Recall@20. On ImageQ, they retained a certain level of performance without our method, achieving Recall@20 ranging from 40-50% on MMQA and 10-20% on WebQA. For E5-V, this can be attributed to its strong capability for understanding textual information through its MLLM backbone, as well as its architecture that converts images into text. While the architecture and training details

of Cohere Embed 3 are not publicly available, its performance suggests that it may adopt a similar architecture or training process to models like E5-V. When our standardization is applied to these models, further improvements are observed on ImageQ; however, it also results in a large drop in TextQ accuracy compared to CLIP-based models and BLIP. This indicates that the benefit of our method is limited when the modality gap is already small.

6.2 Severe Impact of the Modality Gap

To examine how the modality gap affects retrieval performance, we evaluated Recall at various cutoff values of retrieval on ImageQ. Table 4 reports Recall@{1, 5, 20, 100} for each model and dataset.

For Cos, increasing the number of retrieved candidates had almost no effect—Recall@k remained around zero even with $k = 100$. This result clearly indicates that the modality gap severely degrades retrieval performance on ImageQ.

In contrast, our method yields substantial improvements in Recall@k across all tested values of k, demonstrating its effectiveness in bridging the modality gap.

6.3 Pseudo Pairs vs. Manually Labeled Pairs

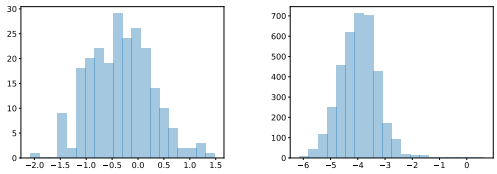
To assess how pseudo pairs affect retrieval, we compared retrieval performances of the Std method and our method. Table 3 shows that the results of our method were equal or higher than those of the Std method. This result demonstrates that pseudo pairs can serve as an effective substitute for manually labeled pairs.

7 Analysis of the Modality Gap

7.1 The Effect of Standardization

To investigate how our method reduced the negative impact of the modality gap, we analyze the distribution of standardized similarity scores on ImageQ. For each ImageQ, we compute the difference between the average standardized similarity scores for image and text candidates in the retrieval database (image mean minus text mean). The distributions on MMQA and WebQA are shown in Figure 4, focusing on CLIP (ViT-B/32) as a representative model that exhibits a clear modality gap.

In MMQA, the distribution is centered slightly below zero, indicating that text scores remain somewhat higher than image scores on average, even after the standardization. In WebQA, the distribution is concentrated mostly on the negative side (around -4), indicating that text candidates are consistently scored higher than images. From these results, we confirm that our method does not fully eliminate the modality gap.



(a) MMQA (b) WebQA

Figure 4: Distributions of the difference between the average standardized similarity scores of image and text candidates across ImageQ queries in the training split of MMQA and WebQA, where the difference is computed as image minus text.

Nevertheless, retrieval performance improves significantly as shown in Section 6. We attribute this to differences in the shape of the cosine similarity score distributions across modalities. Table 5 shows the skewness values in the distributions of similarity scores. CLIP-based models consistently produced more positively skewed similarity distributions for image candidates compared to text candidates. This suggests that some images receive totally higher similarity scores than others in the image database. Such outliers—which often include the correct images—were amplified by our method, allowing them to receive a higher standardized score than most text candidates.

We hypothesized that the skewness in the image similarity distribution stems from the training objective of CLIP-based models These models learn to align images with their paired text, but they are not explicitly trained to capture similarities between texts or between images themselves. As a result, these models yield high similarities to a few image candidates, resulting in a long-tailed distribution. This skewed distribution might align well with our standardization approach, as it amplifies the scores of outliers which often include relevant images.

7.2 Modality Gap in VLMs

We analyze the modality gap in VLMs by investigating both the structure of the embedding space and the distribution of similarity scores.

Following Liang et al. (2022), we apply singular value decomposition (SVD) to project the embeddings of ImageQ queries and their positive examples into a two-dimensional space for visualiza-

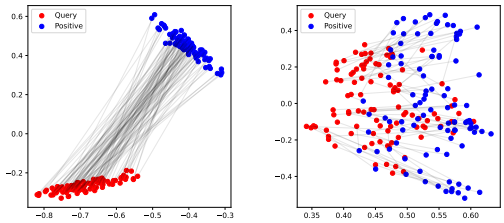
Model	Method	MMQA						WebQA					
		TextQ			ImageQ			TextQ			ImageQ		
		Recall	MRR	NDCG	Recall	MRR	NDCG	Recall	MRR	NDCG	Recall	MRR	NDCG
CLIP (ViT-B/32)	Cos	31.90	26.62	23.90	0.00	0.00	0.00	28.89	21.14	18.89	0.00	0.00	0.00
	Std	31.55	25.78	23.25	52.61	36.65	40.32	23.96	16.38	15.01	32.82	15.20	18.02
	Ours	27.46	18.44	17.88	66.09	45.03	49.86	27.14	19.07	17.29	28.14	13.79	15.98
CLIP (ViT-L/14)	Cos	35.51	28.63	25.86	1.30	0.41	0.62	32.60	24.05	21.45	0.04	0.00	0.01
	Std	35.37	27.60	25.16	62.17	43.32	47.70	28.84	19.37	17.91	43.55	22.54	25.76
	Ours	31.28	21.48	20.54	76.52	58.88	63.05	31.27	21.44	19.69	37.10	20.37	22.75
Long-CLIP-B	Cos	58.67	45.11	43.02	0.00	0.00	0.00	43.93	30.92	28.56	0.00	0.00	0.00
	Std	54.65	40.73	38.76	66.09	47.67	51.94	34.94	23.79	22.05	33.01	14.92	17.98
	Ours	53.33	35.48	35.04	66.96	50.72	54.51	40.44	27.72	25.79	28.59	13.43	15.97
Long-CLIP-L	Cos	63.04	45.56	44.32	0.43	0.11	0.19	45.18	30.36	28.54	0.00	0.00	0.00
	Std	58.39	41.94	40.70	71.74	49.38	54.52	35.66	23.57	22.14	39.84	20.22	23.36
	Ours	57.07	38.60	38.19	73.91	54.04	58.63	41.46	27.14	25.69	35.34	18.38	21.09
BLIP	Cos	41.75	30.20	28.64	0.00	0.00	0.00	37.15	27.07	24.23	0.00	0.00	0.00
	Std	40.92	28.58	27.42	39.57	23.97	27.54	24.00	14.75	14.04	17.62	8.24	9.73
	Ours	36.75	23.33	23.31	43.48	27.45	31.15	31.40	20.71	19.23	14.04	6.35	7.62
Cohere Embed 3	Cos	87.17	78.81	74.72	50.43	20.79	27.61	76.52	59.19	55.86	20.43	8.00	10.16
	Std	72.19	66.33	60.63	52.17	27.24	32.92	54.78	41.69	38.19	27.42	12.36	14.83
	Ours	73.99	63.25	59.20	52.17	28.17	33.61	69.23	52.67	49.36	25.39	11.48	13.73
E5-V	Cos	84.88	66.67	67.20	38.70	17.34	22.06	74.37	54.88	52.27	11.89	5.19	6.37
	Std	80.79	63.33	63.56	41.74	21.33	25.91	48.61	35.04	33.11	21.05	9.75	11.50
	Ours	70.39	53.15	53.12	41.74	21.55	26.09	65.73	48.76	46.11	18.78	8.87	8.87

Table 3: Overall retrieval results on MMQA and WebQA. Recall@20, MRR@20, and NDCG@20 are reported. Cos uses cosine similarity as the retrieval score. Std-L and Std-P apply similarity standardization using modality-specific mean and variance estimated from labeled and pseudo pairs, respectively.

tion. Figure 5 shows the results for CLIP (ViT-B/32) and E5-V. The visualizations of other models and datasets are shown in Appendix D. CLIP exhibits a clear separation between textual queries and positive image items in the embedding space. In contrast, E5-V shows a much smaller gap, suggesting that modality conversion reduces representational disparity between text and images.

We then analyze the cosine similarity scores between queries in the training split of MMQA and their positive examples (either text or image) for CLIP (ViT-B/32) and E5-V. Figure 6 presents the distributions of these scores, separated by the modality of the positive examples. The distributions of other models are shown in Appendix E. As expected, CLIP assigns significantly higher similarities to text examples. E5-V reduces this gap to some extent, but a consistent score difference remains: image positives still tend to receive lower similarity scores than text counterparts.

These results indicate that image captioning reduces modality differences, but does not fully avoid the gap of VLMs. One possible reason is that converting images into textual representations leads to loss of visual information necessary for questions that are difficult to express in language, such as the spatial relationships between objects and the background color. This missing information reduces similarities between queries and relevant candidates compared to text data. Our method avoids this shortcoming. By directly processing image features without converting them into text, our method outperformed E5-V in ImageQ.



(a) CLIP (ViT-B/32) (b) E5-V

Figure 5: 2D visualizations of the embeddings of ImageQ queries in the MMQA training split (blue dots) and their corresponding images (red dots) using SVD. Figures 5a and 5b show the results of CLIP (ViT-B/32) and E5-V, respectively.

8 Conclusion

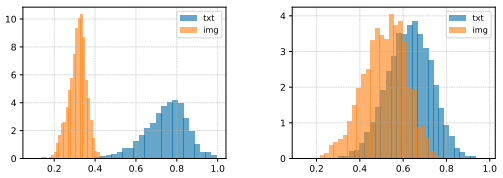
We presented a method for improving multi-modal retrieval by bridging the modality gap without human-created data. Our approach standardizes similarity scores in a modality-specific manner, making them more comparable across modalities. Importantly, it does not require any labeled data or image captions, as it relies on pseudo-positive examples derived from unlabeled queries. Through

Model	Method	MMQA				WebQA			
		1	5	20	100	1	5	20	100
CLIP (ViT-B/32)	Cos	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Ours	37.39	53.48	66.09	72.17	8.16	16.11	28.14	44.78
CLIP (ViT-L/14)	Cos	0.00	0.87	1.30	3.04	0.00	0.00	0.04	0.06
	Ours	50.87	69.57	76.52	81.74	12.90	24.39	37.10	54.76
Long-CLIP-B	Cos	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Ours	43.91	60.43	66.96	76.96	7.89	16.21	28.59	45.46
Long-CLIP-L	Cos	0.00	0.43	0.43	0.87	0.00	0.00	0.00	0.00
	Ours	46.09	63.48	73.91	80.87	11.95	21.39	35.34	52.77
BLIP	Cos	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Ours	21.30	35.22	43.48	56.09	3.78	7.39	14.04	26.66
Cohere Embed 3	Cos	10.00	34.78	50.43	64.78	4.38	9.14	20.43	40.60
	Ours	20.43	38.26	52.17	65.65	6.41	13.52	25.39	43.83
E5-V	Cos	12.17	23.91	38.70	59.57	2.95	6.35	11.89	26.52
	Ours	16.09	28.26	41.74	63.04	5.10	10.49	18.78	36.80

Table 4: Results of Recall@k ($k = \{1, 5, 20, 100\}$) for each model on ImageQ queries in MMQA and WebQA datasets.

Model	MMQA		WebQA	
	Text	Image	Text	Image
CLIP (ViT-B/32)	−0.81	0.21	−1.05	0.45
CLIP (ViT-L/14)	−0.41	0.31	−0.51	0.33
Long-CLIP-B	−1.40	0.35	−1.33	0.59
Long-CLIP-L	−1.88	0.47	−1.30	0.72
BLIP	0.32	0.37	0.59	0.45
Cohere Embed 3	0.26	0.16	0.54	0.25
E5-V	0.81	0.90	0.91	0.76

Table 5: Average skewnesses of cosine similarity distributions for ImageQ queries in the training split of MMQA and WebQA. Each skewness is computed between a query and all candidates in the text or image database, then averaged across all queries per modality.



(a) CLIP (ViT-B/32) (b) E5-V

Figure 6: Distributions of cosine similarity scores between a query and its corresponding positive example (either text or image). The distributions are separated by the modality of the positive example. Figures 6a and 6b show the results of CLIP (ViT-B/32) and E5-V, respectively.

experiments on two multi-modal QA datasets and seven vision-language models, we demonstrated that our method consistently improves image retrieval performance, particularly in scenarios where existing models struggle due to the modality gap. Furthermore, we showed that pseudo-positive examples are sufficient for estimating modality-specific statistics, achieving performance on par with manually labeled data. Our findings highlight the importance of preserving modality-specific information and calibrating similarity scores, rather than relying solely on modality conversion.

Limitations

Our method computes modality-specific similarity statistics from pre-collected datasets and uses them to standardize all similarity scores across modalities. However, this approach assumes that similarity distributions remain stable over time. In real-world systems, new data is constantly being added to databases. Due to new content, these precomputed statistics may become obsolete, leading to suboptimal standardization. Future work should focus on developing mechanisms to dynamically update these statistics.

References

Yinghsan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2021. WebQA: Multihop and Multimodal QA.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural*

Language Processing, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *Preprint*, arXiv:2401.08281.

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

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

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Mei Kobayashi and Koichi Takeda. 2000. Information retrieval on the web. *ACM computing surveys (CSUR)*, 32(2):144–173.

Oleksandr Kolomiyets and Marie-Francine Moens. 2011. A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434.

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

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 *NeurIPS*.

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.

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

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. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Luke Ross, Nils Reimers, Leila Chan Currie, and Elliott Choi. 2024. Introducing multimodal embed 3: Powering ai search. Cohere Blog. Blog post.

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

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. MultiModalQA: complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024a. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer.

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

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024a. VISTA: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3185–3200, Bangkok, Thailand. Association for Computational Linguistics.

Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu.

2024b. MARVEL: Unlocking the multi-modal capability of dense retrieval via visual module plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14608–14624, Bangkok, Thailand. Association for Computational Linguistics.

A Computational Resources

We used two NVIDIA Quadro RTX 6000 GPUs for generating embeddings with E5-V, while only one GPU was used for all other pre-trained VLMs. All retrieval and evaluation experiments were conducted using Faiss (Douze et al., 2024) on CPU only.

B Model List

We evaluated seven pre-trained VLMs in our experiments. Six of them are publicly available on Hugging Face and were accessed as downloadable checkpoints:

- https://huggingface.co/openai/clip-vit-base-patch32
- https://huggingface.co/openai/clip-vit-large-patch14
- https://huggingface.co/BeichenZhang/LongCLIP-B
- https://huggingface.co/BeichenZhang/LongCLIP-L
- https://huggingface.co/Salesforce/blip-itm-base-coco
- https://huggingface.co/royokong/e5-v

We used the Cohere Embed 3 English model (cohere.embed-english-v3) via Amazon Bedrock API in the us-west-2 region.

C Modality-Specific Mean and Variance

Table 6 lists the modality-specific mean and standard deviation for similarity standardization that were used for standardization in our experiments.

D 2D Visualizations of Embeddings

Figures 7–13 illustrate 2D visualizations of embeddings of textual queries (from the training sets of MMQA and WebQA) and their positive examples using singular value decomposition[1].

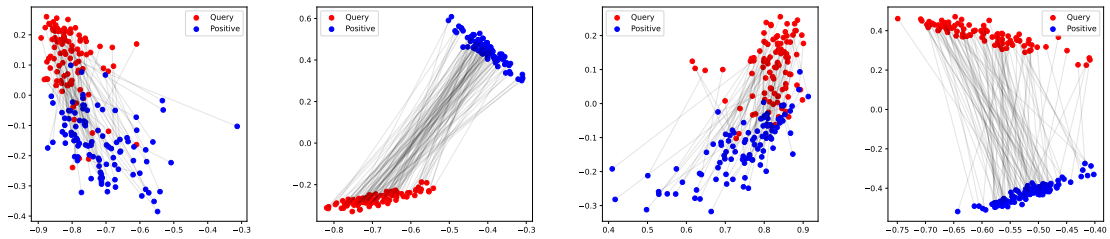
[1]Our visualization code is adapted from https://github.com/Weixin-Liang/Modality-Gap/blob/main/Figure_1_Modality_Gap/visualize.ipynb

E Distributions of Cosine Similarity Scores between Positive Pairs across Modalities

Figure 14 presents the distributions of cosine similarity scores between textual queries (from the training sets of MMQA and WebQA) and their positive examples, separated by the modality of positive examples.

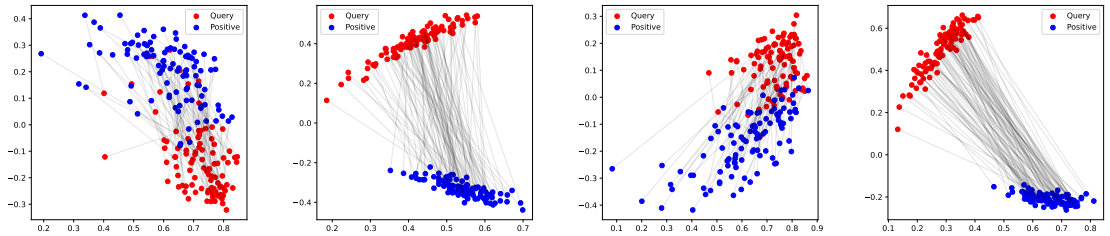
Model	Method	MMQA				WebQA			
		Text		Image		Text		Image	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std
CLIP (ViT-B/32)	Std	0.744	0.105	0.314	0.043	0.789	0.093	0.304	0.035
	Ours	0.841	0.058	0.315	0.023	0.833	0.063	0.335	0.019
CLIP (ViT-L/14)	Std	0.642	0.136	0.280	0.049	0.700	0.122	0.269	0.040
	Ours	0.755	0.088	0.271	0.029	0.749	0.093	0.297	0.023
Long-CLIP-B	Std	0.879	0.050	0.315	0.031	0.895	0.040	0.307	0.024
	Ours	0.898	0.043	0.311	0.017	0.901	0.037	0.324	0.016
Long-CLIP-L	Std	0.828	0.068	0.279	0.048	0.856	0.057	0.258	0.037
	Ours	0.845	0.073	0.264	0.029	0.860	0.059	0.277	0.026
BLIP	Std	0.700	0.116	0.438	0.072	0.724	0.100	0.418	0.059
	Ours	0.791	0.070	0.460	0.038	0.806	0.058	0.489	0.034
Cohere Embed 3	Std	0.629	0.121	0.508	0.082	0.581	0.114	0.490	0.066
	Ours	0.660	0.105	0.512	0.047	0.615	0.082	0.541	0.044
E5-V	Std	0.628	0.102	0.514	0.099	0.635	0.105	0.467	0.084
	Ours	0.649	0.095	0.469	0.085	0.640	0.093	0.534	0.073

Table 6: Modality-specific mean and standard deviation used for standardization during evaluation on MMQA and WebQA datasets. Values are computed separately for text and image modalities, either from labeled or pseudo pairs.



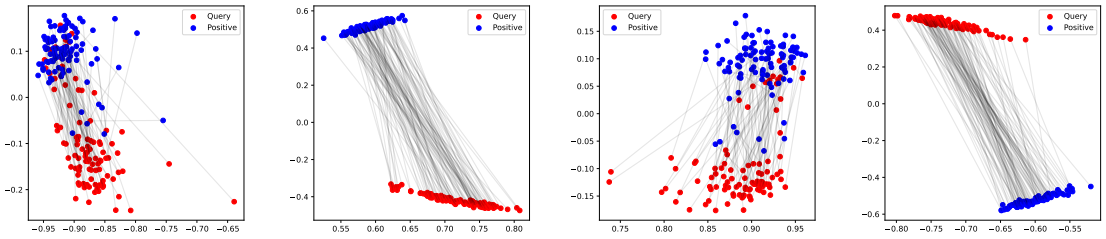
(a) MMQA TextQ (b) MMQA ImageQ (c) WebQA TextQ (d) WebQA ImageQ

Figure 7: 2D visualizations of embeddings from CLIP (ViT-B/32).



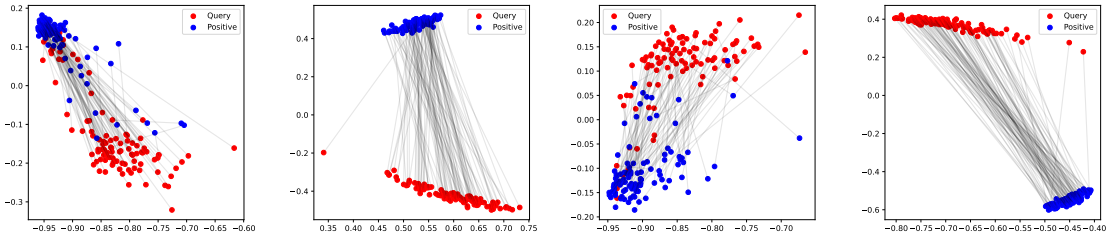
(a) MMQA TextQ (b) MMQA ImageQ (c) WebQA TextQ (d) WebQA ImageQ

Figure 8: 2D visualizations of embeddings from CLIP (ViT-L/14).



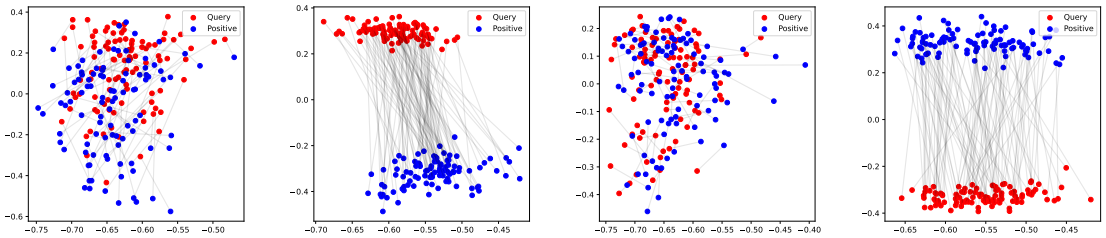
(a) MMQA TextQ (b) MMQA ImageQ (c) WebQA TextQ (d) WebQA ImageQ

Figure 9: 2D visualizations of embeddings from Long-CLIP-B.



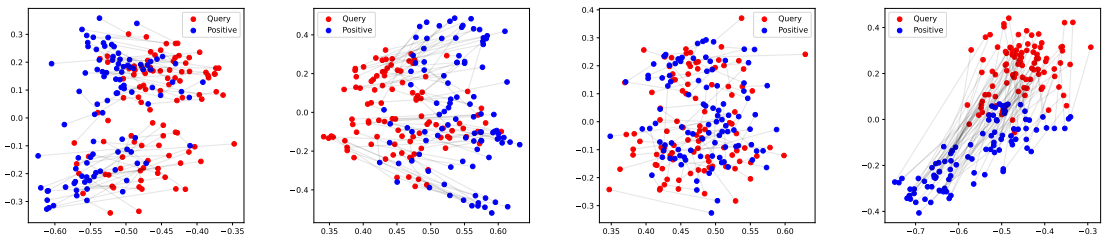
(a) MMQA TextQ (b) MMQA ImageQ (c) WebQA TextQ (d) WebQA ImageQ

Figure 10: 2D visualizations of embeddings from Long-CLIP-L.



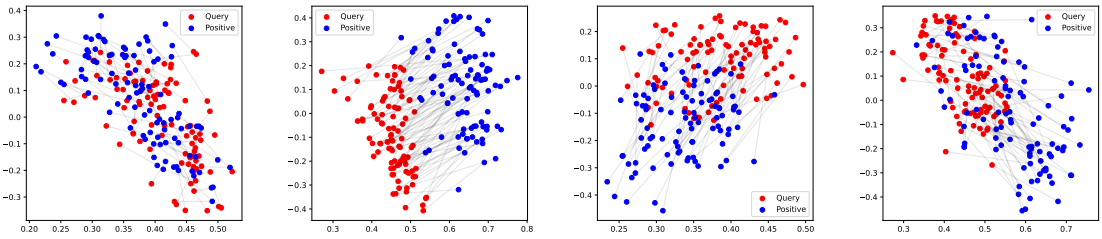
(a) MMQA TextQ (b) MMQA ImageQ (c) WebQA TextQ (d) WebQA ImageQ

Figure 11: 2D visualizations of embeddings from BLIP.



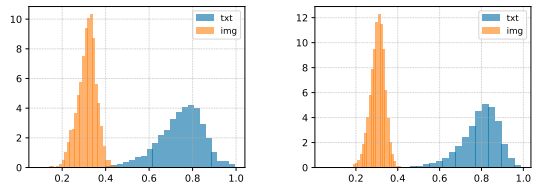
(a) MMQA TextQ (b) MMQA ImageQ (c) WebQA TextQ (d) WebQA ImageQ

Figure 12: 2D visualizations of embeddings from E5-V.

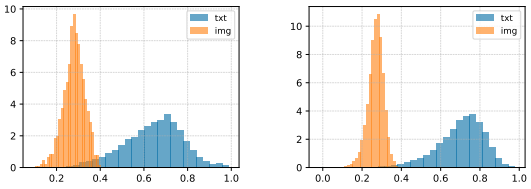


(a) MMQA TextQ (b) MMQA ImageQ (c) WebQA TextQ (d) WebQA ImageQ

Figure 13: 2D visualizations of embeddings from Cohere Embed 3 English.



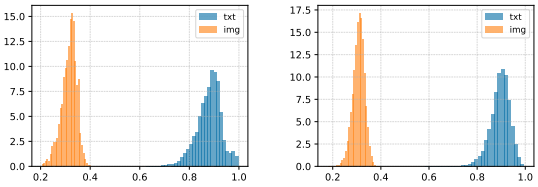
(a) MMQA (b) WebQA



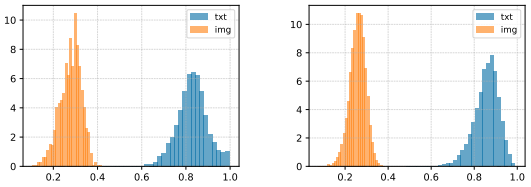
(c) MMQA (d) WebQA

(A) CLIP (ViT-B/32)

(B) CLIP (ViT-L/14)



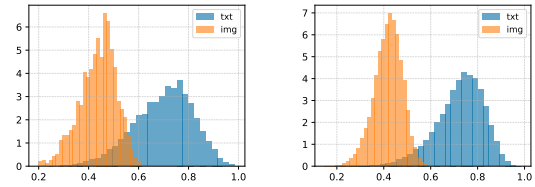
(e) MMQA (f) WebQA



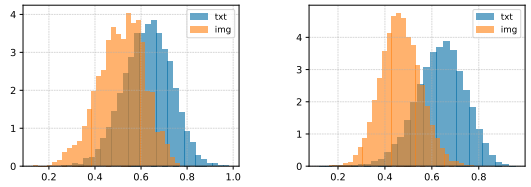
(g) MMQA (h) WebQA

(C) Long-CLIP-B

(D) Long-CLIP-L



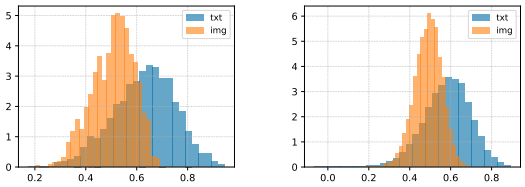
(i) MMQA (j) WebQA



(k) MMQA (l) WebQA

(E) BLIP

(F) E5-V



(m) MMQA (n) WebQA

(G) Cohere Embed 3 English

Figure 14: Distributions of cosine similarity scores between textual queries in the training split of each dataset and their corresponding examples (either text or image).