# MiMIC: Mitigating Visual Modality Collapse in Universal Multimodal Retrieval While Avoiding Semantic Misalignment

Anonymous ACL submission

## Abstract

Universal Multimodal Retrieval (UMR) aims to map different modalities (e.g., visual and textual) into a shared embedding space for multi-modal retrieval. Existing UMR methods can be broadly divided into two categories: *early-fusion approaches*, such as Marvel, which projects visual features into the language model (LM) space for integrating with text modality, and *late-fusion approaches*, such as UniVL-DR, encode visual and textual inputs using separate encoders and obtain fused embeddings through addition. Our pilot study reveals that Marvel exhibits *visual modality collapse*, which is characterized by the model's tendency to disregard visual features while depending excessively on textual cues. In contrast, although UniVL-DR is less affected by this issue, it is more susceptible to *semantic misalignment*, where semantically related content is positioned far apart in the embedding space. To address these challenges, we propose **MiMIC**, which introduces two key innovations: (1) a fusion-in-decoder architecture for effective multimodal integration, and (2) robust training through single-modality mix-in and random caption dropout. Experiments on the WebQA+ and EVQA+ datasets—where image in documents or queries might lack captions—indicate that MiMIC consistently outperforms both early- and late-fusion baselines.

## 1 Introduction

Universal Multimodal Retrieval (UMR) aims to map diverse modalities—such as images and text—into a unified, shared embedding space to facilitate seamless multi-modal search and retrieval (Zhou et al., 2024b; Liu et al.; Zhang et al., 2025b). The necessity of UMR arises from the increasingly heterogeneous nature of digital information; users frequently seek answers that require data from diverse modal sources, where a query in one modality must effectively retrieve relevant content in another. In addition to cross-modal retrieval, both queries and documents in UMR can be multimodal themselves, such as a query composed of an image paired with a textual question.

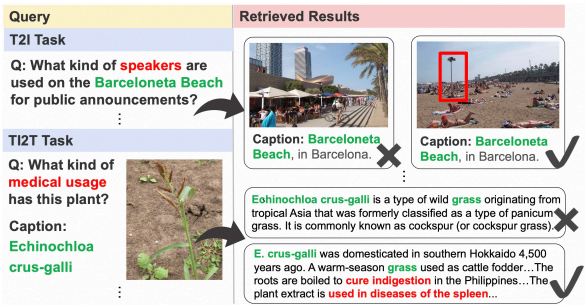



Figure 1: First row demonstrates the visual modality collapse in image document. Second row shows the semantic misalignment, that limits the model to effectively combine and understanding information.

Existing UMR methods can be broadly categorized into two types based on their fusion strategy. Early-fusion approaches project visual features into the language model (LM) space to facilitate multimodal interactions via self-attention mechanisms. Representative methods in this category include Marvel (Zhou et al., 2024b), VISTA (Zhou et al., 2024a), and GME (Zhang et al., 2025b). Conversely, late-fusion approaches—such as UniVL-DR (Liu et al.)—encode visual and textual inputs using separate encoders and derive fused representations through addition.

While early-fusion methods generally outperform late-fusion alternatives, our pilot study with Marvel reveals that this approach exhibits visual modality collapse. This phenomenon is characterized by the model's tendency to disregard visual features in favor of an excessive reliance on textual cues—an observation consistent with recent studies highlighting the "text dominance" issue in Visual Language Models (VLM) and Multimodal LLMs (MLLMs) (Wu et al., 2025; Zheng et al., 2025). In contrast, while late-fusion models like UniVL-DR

are less prone to modality collapse, they are more susceptible to semantic misalignment, where the position of a document in semantic space is heavily influenced by its modality. Figure 1 illustrates these challenges: the first row depicts modality collapse, where ignored visual information regarding the *"speaker"* leads to incorrect retrieval, while the second row demonstrates the model's limited capacity to fully understand the query due to semantic misalignment.

To address these challenges, we propose MiMIC, which introduces two key innovations: (1) a fusion-in-decoder architecture for effective multimodal integration, and (2) a robust training strategy incorporating single-modality mix-in and random caption dropout. Our method employs separate encoders for different modalities and utilizes cross-attention of the LM decoder to selectively aggregate predictive information from multiple modalities into the fused embedding (fusion-in-decoder). During training, we explicitly maintain and "mix in" single-modality representations with the fused embeddings. This design is motivated by the insight that different modalities generalize at different rates (Chaudhuri et al., 2025; Wang et al., 2020). By preserving these individual modality signals, we prevent the model from discarding critical modality-specific information. Furthermore, our caption dropout strategy also forces the model to optimize visual embeddings, preventing over-reliance on textual features. To the best of our knowledge, this is the first work to investigate the modality collapse issue in UMR.

To evaluate our method, we extended the WebQA and EVQA datasets to WebQA+ and EVQA+ to include scenarios with missing modalities, requiring the model to understand information from each modality without over-relying on any one modality. Our experimental results demonstrate that MiMIC consistently outperforms early-fusion baselines, such as Marvel and VISTA, as well as the late-fusion baseline, UniVL-DR. Furthermore, our ablation study validates the critical roles of both the fusion-in-decoder architecture and our robust training strategy in maintaining balanced performance across different retrieval settings.

Our contributions are summarized as follows:

- We conduct an empirical study revealing that existing UMR methods suffer from two distinct failure modes: visual modality collapse in early-fusion models and semantic misalignment in late-fusion models.
- We propose the usage of Fusion-in-Decoder (FID) for multi-modal fusion, which utilizes a Language Model (LM) and cross-attention to dynamically aggregate information from separate encoders.
- We propose a robust strategy consisting of single-modality mix-in to preserve modality-specific signals and random caption dropout to force the model to utilize visual features, effectively mitigating text dominance.
- We demonstrate through extensive experiments that MiMIC consistently outperforms competitive early- and late-fusion baselines, particularly in "imperfect" data settings.

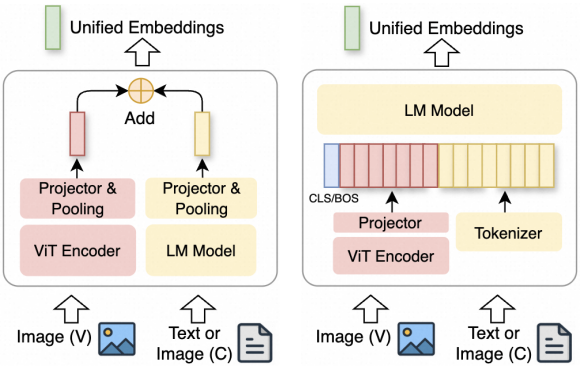
## 2 Related Work

### 2.1 Universial Multi-modal Retrieval

Multi-modal retrieval has attracted growing attention in recent years, with representative benchmarks such as OKVQA (Marino et al., 2019), WebQA (Chang et al., 2022), ViquAE (Lerner et al., 2022), Remuq (Luo et al., 2023), and EVQA (Mensink et al., 2023). Unlike conventional cross-modal retrieval, multi-modal retrieval demands effective representation of compositional information across modalities.

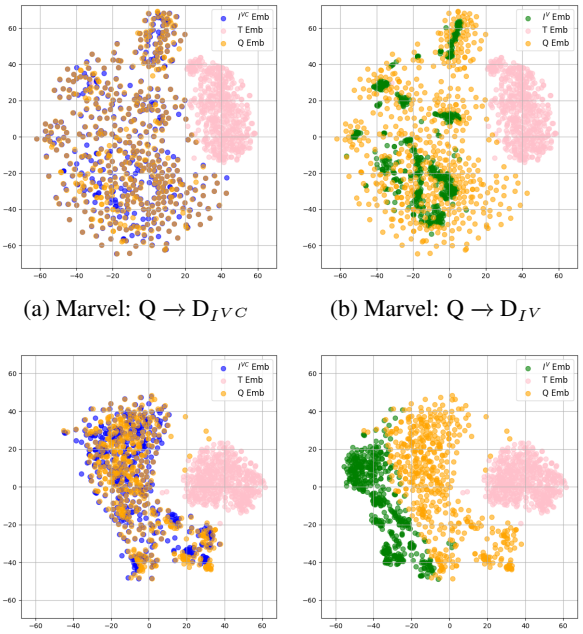
Traditional multimodal retrieval frameworks use a pipeline, performing modality-specific retrieval with separate models before post-ranking. Text retrieval uses models like DPR (Karpukhin et al., 2020), BM25 (Robertson et al., 2009) or BGE (Xiao et al., 2023), while image retrieval relies on CLIP(Radford et al., 2021), BLIP (Chen et al., 2023b) or SIGLIP (Zhai et al., 2023). This separation hinders effective cross-modal integration.

UMR maps multiple modalities into a shared semantic space for intra-, cross-, and multi-modal retrieval. Fusion is key to cross-modal integration, with models mainly using early or late fusion (Figure 2). Early-fusion methods, like Marvel, VISTA, and GME, jointly process textual and visual tokens via self-attention for richer interactions. Late-fusion methods, such as UniVL-DR, rely on dual-tower VLMs (e.g., CLIP, EVA-CLIP, BLIP, SIGLIP) that encode each modality separately and fuse embeddings by addition. Unfortunately, UMR suffers from modality collapse and misalignment issue, which has not been fully addressed in the current liturature.




(a) Late Fusion. (b) Early Fusion.

Figure 2: Different Fusion Strategies. (a) Late Fusion. Use separate image and text encoder (eg. UniVL-DR). (b) Early Fusion. Concatenate image and text embeddings together, then fed into LM. (eg. Marvel, VISTA)



(a) Marvel: Q → $D_{IVC}$ (b) Marvel: Q → $D_{IV}$

(c) UniVL-DR: Q → $D_{IVC}$ (d) UniVL-DR: Q → $D_{IV}$

Figure 3: Embeddings of T2I Queries with paired Image Documents Embeddings ($I^{VC}$ or $I^{V}$) and random Text Documents Embeddings ($T$).

### 2.2 Modality Collapse

Modality collapse occurs when multimodal models fail to integrate information across modalities, overfitting to a dominant one while others lose representational capacity. These problems have been observed in multi-modal applications such as classification (Wang et al., 2020), question answering (Sim et al., 2025), clinical multi-modal prediction (Zhang et al., 2022; Wu et al., 2024). In modern Vision-Language and Multimodal LLMs, this often appears as text modality dominance, where visual cues are neglected in favor of linguistic features (Sim et al., 2025; Chaudhuri et al., 2025). This limitation is particularly concerning given that missing modalities are common in real-world scenarios.

Current studies attribute modality collapse to various factors, including multimodal polysemantic collisions (Chaudhuri et al., 2025), dataset bias or model behavior (Sim et al., 2025), and conflicting gradients (Javaloy et al., 2022). Other work highlights Transformers' sensitivity to missing modalities and the task-dependent robustness of fusion strategies (Ma et al., 2022). Despite this progress, modality collapse remains poorly understood in the context of UMR. This paper provides the first investigation addressing it from the fusion design and training strategies perspectives.

## 3 Pilot Observations

### 3.1 Visual Modality Collapse Issue

**Experimental Setup** We trained UniVL-DR and Marvel on the WebQA dataset. Information about the dataset is given in Section 5.1, and the implementation details of UniVL-DR and Marvel are given in the Appendix A.2. We then utilized the resulting models to extract query and image embeddings for th Text-to-Image (T2I) retrieval task. In this setup, the ground-truth documents are images and the corpus exclusively consists of image documents. We evaluated two specific scenarios: retrieval in $\mathcal{D}_{IVC}$, where documents $I^{VC}$ include both visual content and captions, and $\mathcal{D}_{IV}$, where documents contain only visual information. Finally, we employed t-SNE(van der Maaten and Hinton, 2008) to project the embeddings into a two-dimensional space for qualitative analysis, with the results visualized in Figure 3.

**Results and Discussion** As illustrated in Figure 3, the queries consistently cluster around their corresponding image documents across all cases, indicating effective query-document alignment for the T2I retrieval task. In the multimodal scenario where images are paired with captions (Figure 3a 3c), the fused embeddings exhibit a broad distribution throughout the vector space. However, in the absence of captions (Figure 3b, 3d), a clear performance gap emerges: Marvel embeddings collapse into several highly dense clusters, whereas UniVL-DR maintains a broad spread and more diverse representation of the visual data.

To investigate the impact of visual modality collapse on retrieval performance, we evaluated

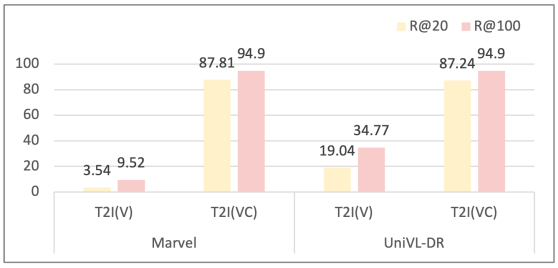
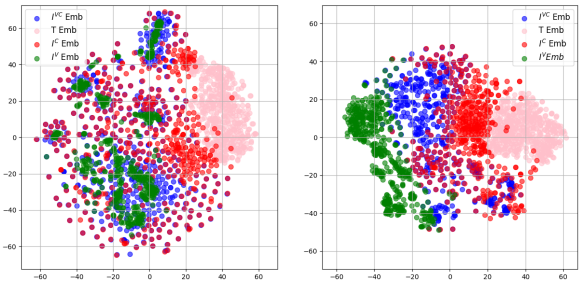



Figure 4: Performance of $T2I^V$ and $T2I^{VC}$ using Marvel and UniVL-DR. Here, $T2I^{VC}$ retrieve in $\mathcal{D}_{I^{VC}}$, a corpus where all images with captions; $T2I^V$ retrieve in $\mathcal{D}_{I^V}$, a corpus where image documents have no caption.



(a) Docs in Marvel space (b) Docs in UniVL-DR space

Figure 5: Embeddings of documents with different modalities, where $I^{VC}$ indicate fused embeddings from both visual and captions.

the Recall@20 (R@20) and Recall@100 (R@100) metrics for both Marvel and UniVL-DR in the T2I task. As shown in Figure 4, when image documents lack captions, Marvel achieves an R@100 of only 9.52%, which is significantly lower compared to the 34.77% achieved by UniVL-DR.

### 3.2 Semantic Misalignment Issue

Although UniVL-DR is less prone to the visual modality collapse issue compared to Marvel, it is more affected by the semantic misalignment problem, where semantically similar content are positioned far apart in the embedding space. We analyze this phenomenon in this section.

**Experimental Setup** To analyze the alignment of document embeddings across different modalities, we utilized the models trained in the Section 3.1 to extract representations for four distinct document types: $I^V$, $I^{VC}$, $I^C$, and $T$. In this context, $T$ refers to textual documents, while $I^V$ and $I^C$ represent the separate visual and caption-based embeddings associated with a single image. Furthermore, $I^{VC}$ denotes the fused multimodal embeddings derived from both visual and textual information. We then employed t-SNE to map the

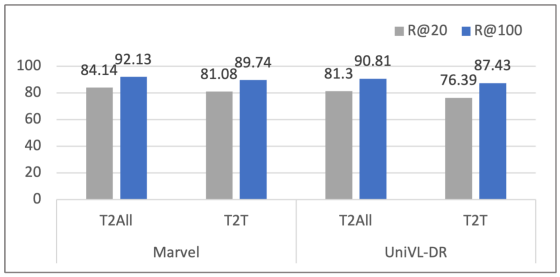



Figure 6: Performance of T2T and T2ALL using Marvel and UniVL-DR. Here, T2ALL retrieve in $\mathcal{D}_{\{I^{VC},T\}}$, a corpus consist of $I^{VC}$ Docs and $T$ Docs, ; T2T retrieve in $\mathcal{D}_T$, a corpus full of $T$ Docs.

| Method | s($I^V$,$T$) | s($I^V$,$I^C$) | s($I^C$,$T$) | s($I^{VC}$,$T$) |
|---|---|---|---|---|
| **Marvel** | 0.103 | 0.494 | 0.206 | 0.098 |
| **UniVL-DR** | 0.073 | 0.270 | **0.491** | **0.391** |

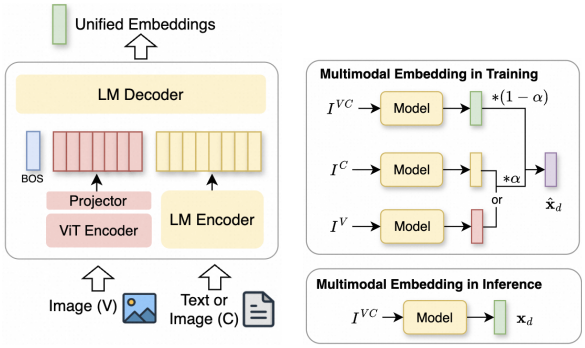
Table 1: Cosine Simlarity between embedding of different modalities of WebQA+ Dataset.

embeddings in a two dimensional space for visualization and showed the scatter plot in Figure 5.

**Results and Discussion** From Figure 5, it is observable that, despite the collapse of the visual modality in the Marvel embedding space, $I^{VC}$ representations remain closely aligned with their corresponding $I^V$ and $I^C$ representations. In contrast, representations of $I^C$, $I^V$ and $I^{VC}$ in the UniVL-DR space tend to cluster separately. There is a gap between different modal representations of the same semantic entity. Table 1 also shows that in UniVL-DR, the average cross-modal similarity s($I^V$,$I^C$) for the same image is 0.27, notably lower than the 0.491 average similarity between image captions and unrelated texts, even though they are merely in same modality. (The furthest T is chosen for calculation for each image). The semantic misalignment issue associated with UniVL-DR leads to its lower overall retrieval performance and T2T performance compared to Marvel on the WebQA dataset, as shown in Figure 6.

## 4 Our Methodology

**MiMIC Architecture** MiMIC uses separate encoders, like late-fusion models, to retain modality-specific features. However, rather than relying on addition to fuse, we introduce cross-attention in a LM decoder to aggregate representations (i.e. Fusion-in-decoder or FID) and enhance semantic alignment to match that of the early-fusion approach. Our architecture is demonstrated in 7a.




(a) FiD Architexture. (b) Single-Modality MixIn.

Figure 7: (left) The encoding of different modalities under our architecture. (right) The multi-modal representation in training and inference phrase.

The framework supports encoder–decoder models (e.g., BART (Lewis et al., 2020), T5, PaLI-X (Chen et al., 2023a), T5Gemma (Zhang et al., 2025a)) and can pair decoder-only LLMs with LLM encoders (e.g., LLM2Vec (BehnamGhader et al., 2024), Qwen3Embedding (Zhang et al., 2025c), LLM2Comp (Zhang et al., 2025d)). Here, we exploit T5 as the LM and CLIP as the vision encoder. Exploration with large VLM-based extensions are left for future work.

More formally, our task is to represent multi-modal documents from a multimodal corpus $\mathcal{D}$ for retrieval, where each document $I_d^{VC} \in \mathcal{D}$ is a multi-modal document with textual and visual modalities in general case. In practice, one of the modalities of the document can be missing, in this case, we denote $T_d$ to indicate full-text document, and $I_d^V$ to indicate the document with only visual information. Similarly, in the query side, we denote $TI_q^{VC}$ for multimodal queries, $T_q$ for full text queries. For a multimodal document, the embedding is obtained by passing both modalities to the MiMIC in Figure 7a. The embedding $\mathbf{x}_d$ is obtained as follows:

$$\begin{aligned} e_d^T &= \text{LM-Encoder}(I^C) \\ e_d^V &= \text{Projector}(\text{Visual-Encoder}(I_d^V)) \\ \mathbf{x}_d &= \text{LM-Decoder}(\text{Concat}(e_d^V, e_d^T)) \end{aligned} \quad (1)$$

where the visual encoder is CLIP visual encoder, LM-Encoder and LM-Decoder indicate T5 encoder and decoder, respectively. For each multi-modal document, we can also obtain the single modal representations: $x_d^V$, visual representation of the $d$ document, and $x_d^T$, the textual representation of the $d$ document, by passing the single modality to both the LM encoder and decoder. Specifically, we have:

$$\begin{aligned} \mathbf{x}_d^T &= \text{LM-Decoder}(e_d^T) \\ \mathbf{x}_d^V &= \text{LM-Decoder}(e_d^V) \end{aligned} \quad (2)$$

For full-text document $T_d$, we obtain the representation $\mathbf{x}_d = \mathbf{x}_d^T$, where $\mathbf{x}_d^T$ is decoded as in Eq. 2. Similarly, the visual-only document has the representation $\mathbf{x}_d = \mathbf{x}_d^V$. We obtain multi-modal query representation $\mathbf{x}_q$ in the same way with encoding documents, using the same network model and parameters.

**Training** Contrastive learning is exploited to train MiMIC so that a positive pair of $(\mathbf{x}_q^i, \mathbf{x}_d^i)$ should be close to each other, whereas a negative pair $(\mathbf{x}_q^i, \mathbf{x}_d^j)$ is far apart, and the loss is shown in Eq. 3. Here, the positive pair means the pair of the (groundtruth) relevant document with the query, and the negative pair is sampled from the set of groundtruth documents of other queries in the batch (in-batch negative). A balance number of samples of different modality $(\mathbf{x}_q, \mathbf{x}_d)$ pairs is supported by the dataset, as shown in Section 5.1.

$$L = -\frac{1}{N} \log \sum_{i=1}^{N} \frac{\exp(f(\mathbf{x}_q^i, \mathbf{x}_d^i))}{\sum_{j=1}^{N} \exp(f(\mathbf{x}_q^i, \mathbf{x}_d^j))} \quad (3)$$

**Single-Modality Mixin** During the training phase, for multimodal documents, we mix the single modality $\mathbf{x}_d^V$ and $\mathbf{x}_d^T$ representations (see Eq.2) into the fused representation $\mathbf{x}_d$, and obtain mix-in representation as follows:

$$\begin{aligned} \hat{\mathbf{x}}_d = &(1-\alpha)\mathbf{x}_d \\ &+ \alpha(\delta \times \mathbf{x}_d^V + (1-\delta) \times \mathbf{x}_d^T) \end{aligned} \quad (4)$$

where $\alpha$ is randomly sampled from $[0, \bar{\alpha}]$ $(\bar{\alpha} < 1)$ to adjust the contribution of single-modal mixin representation, and $\delta$ is also randomly sampled from $\{0, 1\}$ to select which modality to mixin. In the similar way, we can obtain mixin representation for multi-modal queries $\hat{\mathbf{x}}_q$. The single-modal documents and queries are decoded in the same way as previously described. For training with single-modality mixin, we use mixin representations $\hat{\mathbf{x}}_d$, $\hat{\mathbf{x}}_q$ in place of $\mathbf{x}_q$, $\mathbf{x}_d$ in Eq. 3.

**Caption Dropout** Through single-modality mixin, we allow visual encoder to receive more training signals (gradients) through mixin representation. However, one side of these datasets (queries or documents) is still purely text-based,

| Dataset | Query | | | Corpus |
|---|---|---|---|---|
| | Train | Test | Valid | |
| EVQA+ | T / TI<br>15,366 / 32,819 | T / TI<br>500 / 1,049 | T / TI<br>2,455 / 3,750 | T<br>839,692 |
| WebQA+ | T<br>31,766 | T<br>5,000 | T<br>4,966 | I / T<br>389,750 / 787,697 |

Table 2: Dataset Statistics. On the query side, "T" denotes the text modality question and "TI" represents a multimodal question composed of text and image. On the document side, "I" refers to image documents, and "T" refers to text documents. We retain or discard captions of queries or documents including image based on caption ratio.

and text still dominates. So, it is essential that we actively drop captions from images to force the model to rely on visual information for the retrieval. We maintain the *"caption ratio"* as a training hyperparameter.

**ANCE Training** Like Marvel and UniVL-DR, after the first training stage with in-batch negative, we conduct the second training stage, where we adopt hard negative sampling as in ANCE (Xiong et al., 2021). We use the checkpoint obtained from the first stage of training as the retrieval model to encode the multi-modal corpus and retrieve hard negative examples. Besides the hard-negative sampling, both stages exploit single-modal mixin and caption dropout with the same training parameters.

# 5 Experiment

## 5.1 Experimental Setup

**Dataset** WebQA(Chang et al., 2022) is a multi-hop, multi-modal retrieval benchmark, including T2T and T2I retrieval tasks and the number ratio of the two tasks is 1:1. Encyclopedic-VQA(Mensink et al., 2023) is a visual question answering (VQA) dataset, and the questions are designed to be multi-modal. See Appendix A.1 for more details. For our experiments, we extend the WebQA and EVQA datasets to WebQA+ and EVQA+ respectively. The dataset statistics are in Table 2.

In WebQA+, aside from 50% of the images in its corpus lacking captions according to a caption ratio of 0.5, everything else is consistent with WebQA, and the search settings still include T2T and T2I tasks. In EVQA+, this dataset was obtained by sampling the original EVQA dataset and mixing it with WebQA T2T data, the retrieval settings include T2T and TI2T tasks, where the introduction of the T2T task enables the dataset to support a wider range of multimodal retrieval scenarios. In the TI2T setting of EVQA+, each query consists of a question (text) and an image, where the image also preserve captions according to the caption ratio of 0.5.

**Baselines** We compare our results with some baseline methods: UniVL-DR, Marvel, VISTA. Here, CLIP-DPR and Marvel-DPR respectively are UniVL-DRR and Marvel-ANCE without the second stage training (ANCE training with hard negatives). The training details of baselines and MiMIC is given in Appendix A.2,A.3.

**Metrics** We evaluate performance using Recall@1/5/20/100, MRR@10/20, and NDCG@10/20. Recall@k measures the proportion of relevant items retrieved within the top-k results, NDCG@k evaluates ranking quality by accounting for relevance and position, and MRR@k reflects the average reciprocal rank of the first relevant item.

## 5.2 Main Results

Table 3 presents the overall retrieval performance across various multimodal retrieval tasks on WebQA+ and EVQA+. Among the baselines without ANCE training, Marvel-DPR shows a clear advantage over CLIP-DPR on WebQA+ (e.g., +3.89% R@1). On EVQA+, CLIP-DPR proves more robust, outperforming Marvel-DPR by 1.69% in R@1 and 8.79% in R@100. This indicates that EVQA+ places greater emphasis on image understanding, where Marvel is limited due to visual modality collapse.

Among the baselines with ANCE training, Marvel-ANCE is the strongest baseline on WebQA+, surpassing UniVL-DR by 4.03% in R@1. On EVQA+, it remains competitive but lags behind UniVL-DR in R@20 and R@100 (–3.95% in R@100). While Marvel is not good at processing visual information, its semantic space representation is still better than UniVL-DR.

For VISTA, we adopted CLIP-B-32 as the image encoder in order to consistent with other baselines, we denote this variant as VISTA* for reference, while the original VISTA used EVA-CLIP. VISTA* performs poorly on both datasets, showing that it is less robust in modality missing situations.

MiMIC shows superior performance across both evaluated scenarios. In the base configuration without ANCE training, MiMIC outperforms Marvel-DPR, CLIP-DPR, and the VISTA baseline. In particular, in several key metrics, MiMIC even exceeds Marvel-ANCE and UniVL-DR, achieving

| Dataset | Task | Method | R@1 | R@5 | R@20 | R@100 | MRR@10 | NDCG@10 | MRR@20 | NDCG@20 |
|---|---|---|---|---|---|---|---|---|---|---|
| **WebQA+** | T->All | Marvel-DPR | 32.54 | 43.63 | 58.02 | 68.07 | 43.89 | 40.89 | 44.19 | 42.93 |
| | | Marvel-ANCE | 40.84 | 53.13 | 65.33 | 72.10 | 52.61 | 49.48 | 52.78 | 51.00 |
| | | VISTA* | 21.73 | 31.59 | 43.57 | 52.66 | 31.63 | 29.31 | 31.81 | 30.86 |
| | | CLIP-DPR | 28.65 | 39.00 | 54.16 | 66.56 | 39.80 | 36.69 | 40.14 | 38.78 |
| | | UniVL-DR | 36.81 | 49.55 | 62.39 | 69.65 | 49.21 | 46.27 | 49.40 | 47.75 |
| | | MiMIC | 32.62 | 44.35 | 60.29 | 73.07 | 44.62 | 41.75 | 44.96 | 43.88 |
| | | MiMIC-ANCE | 40.96 | 54.88 | 69.67 | 78.87 | 54.11 | 50.97 | 54.38 | 52.77 |
| **EVQA+** | ALL->T | Marvel-DPR | 26.17 | 37.92 | 49.75 | 59.55 | 35.74 | 34.74 | 36.02 | 36.3 |
| | | Marvel-ANCE | 31.85 | 45.26 | 57.51 | 66.8 | 42.26 | 41.5 | 42.5 | 42.96 |
| | | VISTA* | 24.05 | 36.16 | 49.46 | 60.27 | 33.78 | 33.05 | 34.05 | 34.73 |
| | | CLIP-DPR | 27.86 | 41.19 | 55.32 | 68.34 | 38.38 | 37.62 | 38.72 | 39.39 |
| | | UniVL-DR | 30.48 | 45.35 | 58.9 | 70.75 | 41.58 | 41.02 | 41.89 | 42.66 |
| | | MiMIC | 29.41 | 43.08 | 57.06 | 71.13 | 40.03 | 39.30 | 40.36 | 41.15 |
| | | MiMIC-ANCE | 33.36 | 47.90 | 61.28 | 73.87 | 44.21 | 43.64 | 44.51 | 45.28 |

Table 3: Retrieval Performance on WebQA+ (based on a mixed-modal corpus $\mathcal{D}_{\{T,I^V,I^{VC}\}}$) and EVQA+.

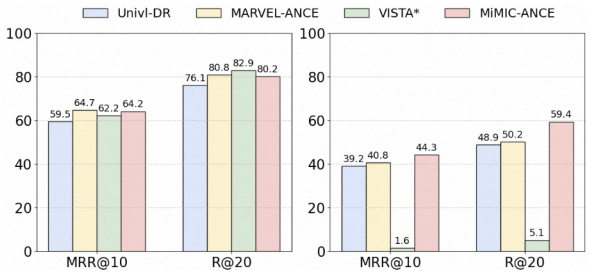



Figure 8: Different modality retrieval tasks: T2T(left) and T2I(right) on WebQA+. All tasks retrieved in a mixed Corpus $\mathcal{D}_{\{T,I^V,I^{VC}\}}$ but statistic separately.

| MiMIC | MRR@10 | Recall@20 | Recall@100 |
|---|---|---|---|
| On WebQA+ Dataset | **44.62** | **60.29** | **73.07** |
| w/o Caption DropOut | 44.05 | 58.05 | 67.77 |
| w/o Single-Modality MixIn | 44.57 | 60.21 | 72.99 |
| w/ Early-Fusion (Marvel-DPR) | 44.22 | 58.59 | 71.84 |
| w/ Late-Fusion (CLIP-DPR) | 37.25 | 54.27 | 71.81 |
| On EVQA+ Dataset | **40.03** | **57.06** | **71.13** |
| w/o Caption DropOut | 38.58 | 52.90 | 63.61 |
| w/o Single-Modality MixIn | 38.19 | 54.57 | 68.51 |
| w/ Early-Fusion (Marvel-DPR) | 38.01 | 54.09 | 68.33 |
| w/ Late-Fusion (CLIP-DPR) | 36.87 | 54.01 | 69.10 |

Table 4: Ablation Study on MiMIC (without ANCE training).

an R@100 of 73.07 on WebQA+, compared to 69.65 for UniVL-DR and 72.10 for Marvel-ANCE. *When integrated with the ANCE strategy, MiMIC-ANCE attains state-of-the-art performance across all categories*. Compared with the strongest baseline, Marvel-ANCE, our MiMIC-ANCE improves R@100 from 72.10 to 78.87 on WebQA+ (a 6.77% absolute gain) and from 66.80 to 73.87 on EVQA+ (a 7.07% absolute gain).

### 5.3 Different Multi-modal Retrieval Settings

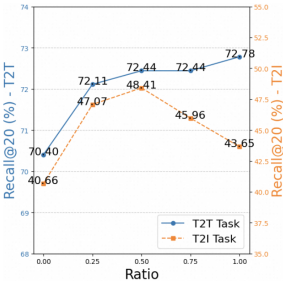
To analyze the contribution of different modalities, we perform evaluation for modality-specific retrieval tasks. For WebQA+, we separately measure T2T and T2I tasks (Figure 8). Note that, the retrieval is over a corpus with mixed documents, which is different from that in Section 3. The results for EVQA+ are reported in Appendix A.4.

The results in Figure 8 show that *MiMIC-ANCE matches strong text retrieval baselines such as Marvel-ANCE and VISTA* in T2T retrieval, while outperforming them substantially in T2I*. In the T2T task, MiMIC-ANCE performs on pair with Marvel-ANCE (R@20 = 80.8) and VISTA* (R@20 = 82.9), achieving R@20 = 80.2. In the T2I task, MiMIC-ANCE achieves a clear advantage, reaching R@20 = 59.4, which surpasses Marvel-ANCE (R@20 = 50.2) by +9.2 and UniVL-DR (R@20 = 48.9) by +10.5. Notably, while VISTA* excels in text retrieval, its performance drops sharply in T2I (R@20 = 5.1), revealing severe modality bias. In contrast, MiMIC-ANCE maintains balanced and robust performance across modalities, consistent with the observations in Table 3.

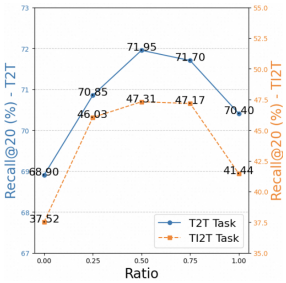
### 5.4 Ablation Study

The ablation study examines the contribution of each component in MiMIC (without ANCE training). We evaluate performance changes under four settings: (1&2) w/o Caption Dropout or w/o Single-Modality Mixin, where each training strategy is removed in turn; (3) w/ Early Fusion (Marvel), using the Marvel-DPR architecture with MiMIC's training strategies without ANCE training; and (4) w/ Late Fusion, using the CLIP-DPR architecture (which is also the same with UniVL-DR) with the same training strategies. The second and third variants assess the effect of the fusion method.

The results in Table 4 indicate that caption

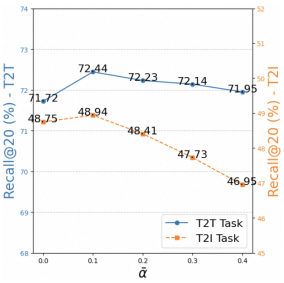



(a) on WebQA+ Datasets

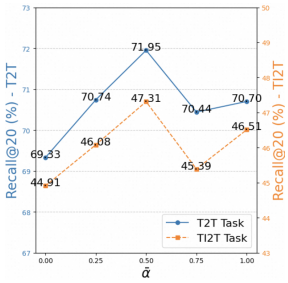



(b) on EVQA+ Datasets

Figure 9: The Recall@20 performance of MiMIC using different Caption Ratios.




(a) on WebQA+ Datasets




(b) on EVQA+ Datasets

Figure 10: The Recall@20 performance vary with the increasing $\bar{\alpha}$ in Single-Modality MixIn of MiMIC.

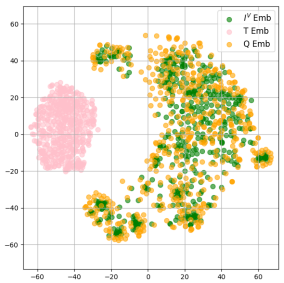
dropout is essential for both datasets, leading to substantial declines in R@20 (-2.24 for WebQA+ and -4.16 for EVQA+) and R@100 (-5.3 for WebQA+ and -7.52 for EVQA+) when removed. In contrast, single-modality mixin has a stronger effect on EVQA+, removing this strategy causes a 1.6-point drop in R@100, compared to only 0.06 on WebQA+.

By comparing MiMIC with variants where the fusion-in-decoder design is replaced by early fusion (as in Marvel-DPR) or late fusion (as in CLIP-DPR), we observe noticeable drops in all metrics across both datasets. These results confirm the robustness of the fusion-in-decoder architecture in maintaining balanced multimodal retrieval performance.

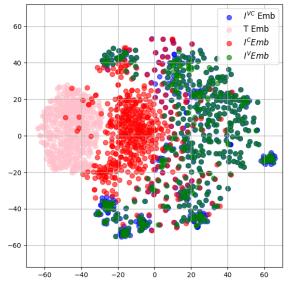
### 5.5 Further Analysis

**The Impact of Caption Ratio** We vary the caption dropout ratio during training and evaluate performance on T2T and T2I tasks in WebQA+, as well as T2T and TI2T tasks in EVQA+. As shown in Figure 9, the best results for T2I and TI2T tasks are achieved with caption ratios between 0.25 and 0.75. Interestingly, both extreme settings—0 (all captions used) and 1.0 (no captions)—degrade performance, suggesting that caption may act as a bridge, helping to map the image to semantic space.

**The Impact of Single-modality Mixin** To examine the effect of modality mixin, we vary the upper bound of the mix-in ratio, $\bar{\alpha}$. The influence of single-modality mix-in appears dataset-dependent: the optimal ratio is around 0.0–0.2 for WebQA+, and approximately 0.5 for EVQA+. For datasets that rely more on vision when understanding semantics, performance improves more obviously as $\bar{\alpha}$ increases, but excessive increases will introduce more noise and lead to performance degradation.




(a) MiMIC: Q $\rightarrow$ $D_{IV}$




(b) Docs in MiMIC space

Figure 11: Embeddings in MiMIC Space

**Embedding Space of MiMIC** Consistent with Section 3, we visualized the representations from MiMIC. As shown in Figure 11a, the query embeddings align effectively with their corresponding image document embeddings $I^V$, even in modality missing situation. And the image representations exhibit a diverse and well-distributed pattern without obvious representation collapse. Figure 11b demonstrates that $I^V$ and $I^{VC}$ exhibit consistent distributions with no significant modality gap. The caption embeddings $I^C$ serve as a semantic bridge: Some captions surround the image, while others are somewhere between the text and image modalities.

## 6 Conclusion

This work identifies and addresses two primary failure modes in UMR: visual modality collapse and semantic misalignment. By introducing Fusion-in-Decoder (FID), we leverage language models and cross-attention to dynamically integrate multimodal signals. Combined with our robust training strategy—incorporating single-modality mix-in and caption dropout—our approach significantly outperforms strong baselines on WebQA+ and EVQA+, demonstrating particular resilience in incomplete data scenarios.

## Limitations

Our work offers pioneering insights into the visual collapse and semantic alignment issues of multimodal representation spaces in UMR. Current research has not yet balanced the performance of various multimodal tasks. Our work is also a preliminary exploration, primarily inspired by the performance of different fusion strategies on the modality missing WebQA+ and EVQA+ datasets. Future research will investigate the impact of dataset in different modality types on shaping the multimodal representation space, and extend our work to larger models to accommodate training and learning with more data. Furthermore, a deeper investigation into dynamically adjusting $\bar{\alpha}$ in Single-Modality MinIn is left for future work, as different training stages may benefit from different mix-in levels. There is still considerable room for research on how multimodal representations can fully utilize information from two modalities and on the competition and cooperation between different modalities in UMR task.

## References


Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. 2025. A closer look at multimodal representation collapse. *arXiv preprint arXiv:2505.22483*.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, and 24 others. 2023a. Pali-x: On scaling up a multilingual vision and language model. *Preprint*, arXiv:2305.18565.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023b. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, Singapore. Association for Computational Linguistics.

Adrián Javaloy, Maryam Meghdadi, and Isabel Valera. 2022. Mitigating modality collapse in multimodal vaes via impartial optimization. In *International Conference on Machine Learning*, pages 9938–9964. PMLR.

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 *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G Moreno, and Jesús Lovón Melgarejo. 2022. ViQuAE, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'22, New York, NY, USA. Association for Computing Machinery.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880.

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

Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2023. End-to-end knowledge retrieval with multi-modal queries. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8573–8589, Toronto, Canada. Association for Computational Linguistics.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18177–18186.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, Andre Araujo, and Vittorio Ferrari. 2023. Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories. In *ICCV*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and

1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. In *Foundations and Trends® in Information Retrieval*.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. 2025. Can vlms actually see and read? a survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. 2025. When language overrules: Revealing text dominance in multimodal large language models. *arXiv preprint arXiv:2508.10552*.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. 2024. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations (ICLR)*.

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

Biao Zhang, Fedor Moiseev, Joshua Ainslie, Paul Suganthan, Min Ma, Surya Bhupatiraju, Fede Lebron, Orhan Firat, Armand Joulin, and Zhe Dong. 2025a. Encoder-decoder gemma: Improving the quality-efficiency trade-off via adaptation. *Preprint*, arXiv:2504.06225.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. 2022. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 2418–2428, New York, NY, USA. Association for Computing Machinery.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025b. Gme: Improving universal multimodal retrieval by multimodal llms. *Preprint*, arXiv:2412.16855.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025c. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Yeqin Zhang, Yizheng Zhao, Chen Hu, Binxing Jiao, Daxin Jiang, Ruihang Miao, and Cam-Tu Nguyen. 2025d. Learning to compress: Unlocking the potential of large language models for text representation. *Preprint*, arXiv:2511.17129.

Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, and 1 others. 2025. Mllms are deeply affected by modality bias. *arXiv preprint arXiv:2505.18657*.

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 Appendix

## A.1 Detail Introduction of Datasets

**WebQA** This dataset contains images and passage snippets that are crawled from the general Web and Wikipedia. It's a multi-modal retrieval that includes both text-to-text (T2T) retrieval, where a text query retrieves text documents, and text-to-image (T2I) retrieval, where a text query retrieves image documents. The ratio of these two types of queries in the WebQA dataset is 1:1. Its corpus is a hybrid corpus containing both images and text, with image documents all having captions. We use the same WebQA dataset as UniVL-DR and Marvel.

**EVQA** Encyclopedic-VQA is a visual question answering (VQA) dataset that requires the ability to understand and reason about detailed encyclopedic knowledge. The dataset is larger in scale, and the questions are designed to be truly multimodal. This means that answers cannot be based solely on images or text alone. Each answer is annotated with supporting evidence from the corresponding Wikipedia section.

We use the processed EVQA datset from *BByrneLab/M2KR* and create an EVQA+ datasets based on EVQA, merging WebQA's T2T task in it and making it an any-modality-to-text-corpus version. The captions of the query-side images in the EVQA+ dataset are taken from the title of its ground truth wiki documentation.

## A.2 Implementation Details of Baseline

**UniVL-DR Training** For training CLIP-DPR, the first stage checkpoint UniVL-DR, we start from the ViT-B/32 version of CLIP and continuously train CLIP on the WebQA+ dataset or EVQA+ dataset with in-batch negatives. Here caption ratio of datasets in training is set to 1, which is equivalent to the original WebQA dataset, to align with the original result. We truncate texts with the max length of 77 and set accumulate step as 1, batch size to 64, max training epoch to 20, and the temperature hyperparameter = 0.01. The learning rate is 5e-6 for the WebQA+ dataset and 1e-5 for the EVQA+ dataset. The cosine annealing strategy is used to schedule the learning rate during training.

In the second stage, for training UniVL-DR, we retrieve Top 100 documents using CLIP-DPR and sample two hard negatives of different modalities (k = 1) from these candidates. All models are tuned with AdamW optimizer, are evaluated per 500 steps, and set early stop step as 5. Training is conducted on a single A100.

| Method | T2T | T2I |
|---|---|---|
| **VISTA*(CLIP-B-32)** | 82.91 | 5.1 |
| **VISTA(EVA-CLIP-16)** | 82.58 | 42.8 |

Table 5: R@20 performance of VISTA on different modality retrieval tasks of WebQA+ Dataset.

**Marvel Training** Marvel model is based on a pre-trained T5 model and uses CLIP's visual encoder as an image tokenizer, and 128 max text token length and 49 image token legnth are set. Following the original setting in paper, we trained on a single A100, using a batch size of 64 and a learning rate of 5e6 for WebQA+ datasets and a batch size of 64 and a learning rate of 1e5 for EVQA+ datasets. Caption ratio of datasets in training is set to 1 to align with the original result. And Temperature hyperparameter = 0.01. And early stopping is implemented like UniVL-DR.

We follow the two-stage training of Marvel: In the first stage, we train the model using only in-batch negative examples. In the second stage, we use both self-mined hard negatives and in-batch negatives to obtain the Marvel-ANCE model, similar to that of UniVL-DR.

**VISTA Training** VISTA model is based on a pre-trained BGE-base model and uses EVA-CLIP's visual encoder as an image tokenizer, and 512 max text token length and 196 image token legnth are set. For a fair comparison, we replaced EVA-CLIP with CLIP-B-32, and named VISTA* in paper. The performance differences between VISTA* and the original VISTA on the WebQA+ dataset are shown in the Table 5. We can see that VISTA is not good at T2I, especially, when the image encoder is replaced by CLIP-B-32.

Following the original setting in paper, we trained on a single A100, using a batch size of 128 and a learning rate of 1e5 for WebQA+ datasets and a batch size of 64 and a learning rate of 1e5 for EVQA+ datasets. Caption ratio of datasets in training is set to 1.

## A.3 Implementation Details of MiMIC

During training MiMIC, we employ the T5 text encoder and CLIP image encoder initialized with the

t5-ance checkpoint from OpenMatch and clip-vit-base-patch32 checkpoint from OpenAI, truncate the text with the max length of 128 and set the batch size to 64, learning rate=$5e-6$, max training epoch to 20, and the temperature hyperparameter $\tau$ = 0.01. And cosine annealing strategy is used to schedule the learning rate during training. All models are tuned with AdamW optimizer, are evaluated per 500 steps, and set early stop step as 5. Like the Implementation of UniVL-DR and Marvel.

We set the caption ratio to 0.5 of WebQA+ and EVQA+ datasets, meaning that only half of the multi-modal candidates or queries in the batch have captions. Additionally, when calculating the training features for the Multimodal Candidates $I_d^{VC}$ from the WebQA+ dataset, we set $\bar{\alpha}$ to 0.1, and when calculating the training features for the Multimodal Queries $TI_q^{VC}$ from the EVQA+ dataset, we set $\bar{\alpha}$ to 0.5.

In our second training stage, we retrieve Top 100 documents using MiMIC and sample one image hard negative (can be have or not have caption) and one text hard negative (k = 1) from these candidates for WebQA+ dataset or sample one text hard negative (k = 1) for EVQA+ dataset. After training using in-batch hard negatives and hard negatives, we get the MiMIC-ANCE model.

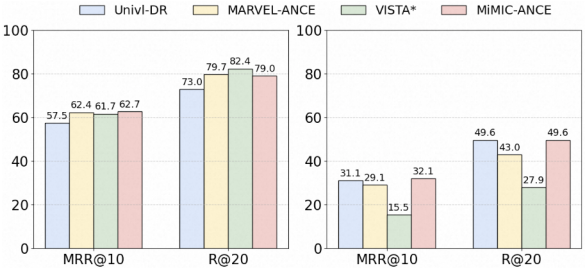



Figure 12: The performance of T2T(left) and TI2T(right) Task on EVQA+ dataset using different methods.

### A.4 Different Multi-modal Retrieval of EVQA+

Here we show the performance of different modality subtask on EVQA+ Dataset in Figure 12. Marvel-ANCE performs consistently and well in the T2T task, but lags behind MiMIC-ANCE in the TI2T task. While Univl-DR has relatively low metrics in the T2T task, it shows good competitiveness in the TI2T task. Similarly, VISTA*, as on the WebQA+ dataset, exhibits significant differences across different tasks. In the TI2T task, MiMIC-ANCE performs best, ranking first in both MRR@10 (32.1) and R@20 (49.6). This indicates that the model has stronger feature modeling capabilities and retrieval accuracy when handling complex visual-text joint queries.