Zero-Shot Multimodal Retrieval with Multi-Scale Contextual Representations

Sourajit Saha and Tejas Gokhale
University of Maryland, Baltimore County
{ssaha2, gokhale}@umbc.edu

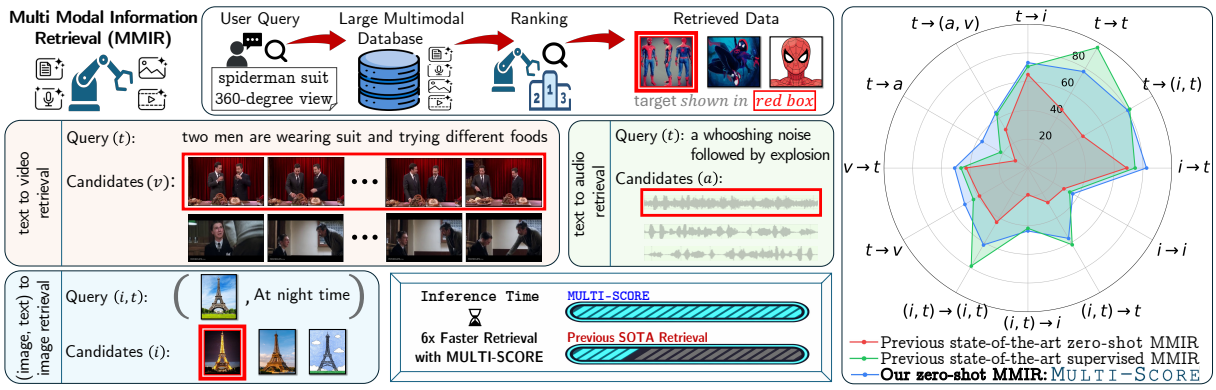

Figure 1: MULTI-SCORE enables zero-shot training-free retrieval and re-ranking across text, image, video, and audio modalities, seamlessly handling unimodal, cross-modal, and composite multimodal query–candidate combinations. This approach leverages multi-scale Matryoshka embeddings for efficient coarse retrieval, followed by multimodal re-ranking using multimodal question answering, token-wise contextual aggregation, and in-context-learning. On 12 MMIR tasks (32 datasets), MULTI-SCORE achieves state-of-the-art zero-shot performance, while matching or surpassing supervised methods on most tasks. Project page: https://sourajitcs.github.io/multiscore/

Abstract

In multimodal information retrieval (MMIR), candidates relevant to an input query need to be retrieved from a database, where the query and database items span different modalities. As real-world databases evolve, repeatedly annotating and indexing data and re-optimizing domain-specific models across modalities is impractical. We present MULTI-SCORE, a fine-tuning-free, two-stage MMIR approach that couples efficient candidate filtering with fine-grained multimodal re-ranking. Stage-1 adopts Matryoshka representations to efficiently filter out low-relevance candidates without expensive similarity computations on full-scale representations for the entire database. Stage-2 re-ranks the filtered candidates by computing their fine-grained multimodal contextual representations with two scoring functions for semantic alignment using chain-of-thought prompting and question-answering. Experiments demonstrate state-of-the-art zero-shot retrieval on 12 MMIR tasks across 32 datasets while outperforming supervised methods on 23 datasets.

1 Introduction

The goal of multimodal information retrieval (MMIR) is to identify and rank database items that best match a user query across modalities such as text, image, video, and audio. For e.g., retrieving a video for the query *"Bill Murray wearing a blue cardigan interviewed by David Letterman"* requires fine-grained vision and language understanding, multimodal reasoning, and cross-modal grounding. While recent approaches have leveraged advances in multimodal large language models (MLLMs) (Team, 2025a; Chu et al., 2024; Xiao et al., 2024) and dual encoders (Radford et al., 2021) for MMIR (Cheng et al., 2021; Shvetsova et al., 2022; Xu et al., 2021), they rely on annotated training data and lack generalization across different modalities. To overcome these challenges, we propose MULTI-SCORE, a fine-tuning-free two-stage retrieval framework: Stage-1 performs efficient candidate filtering and Stage-2 refines the results using fine-grained multimodal alignment, thus enabling scalable zero-shot multimodal retrieval.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20304–20324
July 2-7, 2026

For the first stage of efficient candidate filtering from a large database to a smaller set of candidates, we propose Pyramid Rank, an algorithm that performs efficient retrieval through hierarchical (pyramidal) filtering across embedding resolutions. Pyramid Rank builds upon Matryoshka Representation Learning (MRL) (Kusupati et al., 2022; Zhang et al., 2025a), where embeddings of increasing dimensionality encode progressively finer semantic details. As computing query-candidate similarity for large database is expensive, we derive an admissible similarity upper bound for Pyramid Rank and show how similarity can be computed using much lower-dimensional embeddings. Pyramid Rank starts from low-dimensional embeddings and quickly filters out low-similarity candidates using this upper bound without ever computing similarity with the high-dimensional MRL embeddings for the entire database and progressively refines the search for efficiently ranking the top-K candidates.

For the second stage, to re-rank the candidates from Stage-1, we apply fine-grained multimodal query-candidate alignment via two complementary components, Bidirectional-CoT Embedding Score and Question Answering Relevance Score. Bidirectional-CoT Embedding Score is a bidirectional chain-of-thought (CoT) prompting strategy to infer alignment in both *query-to-candidate* and *candidate-to-query* directions. Using Jiang et al. (2024a)'s one-word limitation principle, we insert a special `<emb>` token, whose preceding hidden state (that aggregates context from the entire prompt) is extracted as the final embedding vector to compute similarity. To compute Question Answering Relevance Score, we use MLLMs to convert each query into question-answer pairs and treat each candidate as potential answer sources. We obtain a score representing a candidate's relevance to the query by measuring how accurately questions can be answered with information found in that candidate. Our key contributions and findings include:

- MULTI-SCORE, a fine-tuning free zero-shot MMIR system that is both efficient and capable of unified alignment across text, image, video, and audio modalities, in unimodal, cross-modal, and composite query–candidate combinations;
- Pyramid Rank, an algorithm for efficient candidate filtering with multi-scale representations;
- Bidirectional-CoT Embedding Score and Question Answering Relevance Score for fine-grained multimodal alignment to re-rank candidates; and
- Comprehensive experiments demonstrating that MULTI-SCORE achieves state-of-the-art zero-shot performance (and often surpassing supervised models) at a lower inference time on (a) 12 MMIR tasks across 32 datasets as well (b) jointly retrieving over these 32 datasets combined with 5.7M multimodal database items.

2 Related Work

Multi-Scale "Pyramidal" Representations have underpinned visual recognition for decades, from hierarchical and progressive vision by Marr (1982); Barrow and Tenenbaum (1981), Gaussian and Laplacian pyramids (Adelson et al., 1984), multi-resolution histograms (Hadjidemetriou et al., 2004), spatial pyramid features (Lazebnik et al., 2006), and others. MRL (Kusupati et al., 2022) produces multi-scale "pyramidal" embeddings that are semantically consistent and support coarse-to-fine retrieval through progressively expandable representation scales (Zhang et al., 2025a).

Multimodal Information Retrieval. MLLMs leverage large scale multimodal pretraining (Team, 2025b,a; Xiao et al., 2024), while dual encoders offer cross-modal alignment (Li et al., 2023; Guzhov et al., 2022; Zhai et al., 2023; Radford et al., 2021; Luo et al., 2023), for MMIR. However these models exhibit modality-specific limitations, as retrieval models trained on text–image pairs often generalize poorly when transferred to other modalities such as retrieving video (Liu et al., 2025). Training such models jointly on retrieval datasets across modalities is computationally expensive and requires large scale annotation. To reduce annotation costs, prior work (Ge et al., 2022; Wang et al., 2022; Cheng et al., 2021; Shvetsova et al., 2022; Xu et al., 2021) proposes fine-tuning on diverse auxiliary tasks, but fine-tuning depends on human-labeled data and identifying which auxiliary tasks maximally benefit retrieval across modalities remains infeasible at scale.

3 Method

Given a query q and a database with N items $\mathcal{C}=\{c_n\}_{n=1}^{N}$, information retrieval seeks to rank the top-K candidates in $\mathcal{C}$ with the highest relevance to q. Queries and candidates are both represented as vector embeddings, and relevance is computed using measures such as cosine similarity between the query and candidate embeddings. Our method unifies multimodal retrieval into an efficient pipeline across modalities.

We present, MULTI-SCORE, short for Multi-Scale Contextual Representations, a fully zero-shot two-stage approach that does not require dataset-specific large-scale multimodal pre-training (Zhang et al., 2024a; Kim et al., 2025; Liu et al., 2025), auxiliary fine-tuning (Ge et al., 2022; Wang et al., 2022), or additional data annotation. Stage-1 performs theoretically guaranteed efficient similarity computations for text-only ranking and filtering to obtain top-K relevant candidates ($K<<N$). Stage-2 computes fine-grained contextual multimodal embeddings via (1) a bidirectional chain-of-thought prompting approach and (2) a question-answering approach to produce refined relevance scores for further re-ranking the top-K candidates retrieved from Stage-1 for state-of-the-art MMIR.

3.1 Stage-1: Efficient Candidate Filtering

In Stage-1, all queries (processed at inference) and database items (pre-processed) are first converted to text via image/video captioning with Qwen3-VL-8B (Team, 2025b), audio transcription using Qwen2-Audio-7B (Chu et al., 2024); and audio/video captioning using Qwen2.5-Omni-7B (Xu et al., 2025a). This gives us text-only queries and database items for coarse ranking at Stage-1. Then, we use Qwen3-MRL (Zhang et al., 2025b) to compute embeddings x_c for all database entries c and x_q for input queries, and normalize them to be unit-norm. Qwen3-MRL feature extractor produces embeddings at L=6 levels, with the smallest length d=32 (level ℓ=1) and largest length D=1024 (level L=6) with following properties:

$$\mathrm{x}_\mathrm{c}^{(\ell)} \in \mathbb{R}^{2^{\ell-1}d} \text{ ... length of level-}\ell\text{ vectors} \quad (1)$$

$$\mathrm{x}_\mathrm{c}^{(\ell)} = \mathrm{x}_\mathrm{c}^{(\mathrm{L})}[1:2^{\ell-1}d] \text{ ... nested MRL embeddings} \quad (2)$$

$$\|\mathrm{x}_\mathrm{c}^{(\mathrm{L})}\|_2 = \|\mathrm{x}_\mathrm{q}^{(\mathrm{L})}\|_2 = 1 \text{ ... unit-norm level-}L\text{ vectors} \quad (3)$$

We create zero-padded candidate and query embeddings $\mathrm{z}_\mathrm{c}^{(\ell)}, \mathrm{z}_\mathrm{q}^{(\ell)} \in \mathbb{R}^D$ as:

$$\mathrm{z}_\mathrm{c}^{(\ell)} = \mathrm{concat}(\mathrm{x}_\mathrm{c}^{(\ell)}, zeros(D-2^{\ell-1}d)); \text{ and}$$
$$\mathrm{z}_\mathrm{q}^{(\ell)} = \mathrm{concat}(\mathrm{x}_\mathrm{q}^{(\ell)}, zeros(D-2^{\ell-1}d)). \quad (4)$$

3.1.1 Deriving an Upper-Bound for Similarity

For retrieval, we ideally want to compute similarity of each candidate vector $\mathrm{x}_\mathrm{c}^{(\mathrm{L})}$ with the query vector $\mathrm{x}_\mathrm{q}^{(\mathrm{L})}$, at the highest (most fine-grained) representation level L. However, this is computationally expensive, with a complexity of $O(2^L N)$. Therefore, Stage-1 seeks to reduce this computational complexity by leveraging multi-scale MRL representations. From lower level ℓ, we derive an upper bound for $\langle \mathrm{x}_\mathrm{q}^{(\mathrm{L})}, \mathrm{x}_\mathrm{c}^{(\mathrm{L})} \rangle$ in terms of lower-dimensional $\mathrm{x}_\mathrm{c}^{(\ell)}, \mathrm{x}_\mathrm{q}^{(\ell)}$, thus significantly reducing compute cost (Fig. 2) for similarity computation.

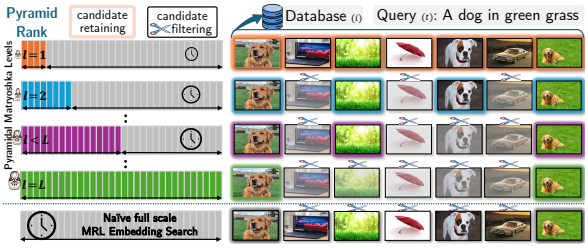

Figure 2: **Stage-1 Pyramid Rank**: with smaller MRL embedding (l=1, 2, ...), Pyramid Rank filters low-similarity candidates early, reducing computations to retain top-K relevant candidates for Stage-2 refinement. This leads to large efficiency gains without loss in accuracy, compared to naïve full-scale MRL search (bottom).

$$\langle \mathrm{x}_\mathrm{q}^{(\mathrm{L})}, \mathrm{x}_\mathrm{c}^{(\mathrm{L})} \rangle = \underbrace{\langle \mathrm{z}_\mathrm{q}^{(\ell)}, \mathrm{z}_\mathrm{c}^{(\ell)} \rangle}_{known} + \underbrace{\langle \mathrm{x}_\mathrm{q}^{(\mathrm{L})} - \mathrm{z}_\mathrm{q}^{(\ell)}, \mathrm{x}_\mathrm{c}^{(\mathrm{L})} - \mathrm{z}_\mathrm{c}^{(\ell)} \rangle}_{unknown\,\text{☹}} \quad (5)$$

Applying[1] the Cauchy-Schwartz inequality (Schwarz, 1890) on the *unknown* term, and given that both $\mathrm{x}_\mathrm{q}^{(\mathrm{L})}$ and $\mathrm{x}_\mathrm{c}^{(\mathrm{L})}$ are unit norm:

$$\langle \mathrm{x}_\mathrm{q}^{(\mathrm{L})}, \mathrm{x}_\mathrm{c}^{(\mathrm{L})} \rangle \leq \langle \mathrm{z}_\mathrm{q}^{(\ell)}, \mathrm{z}_\mathrm{c}^{(\ell)} \rangle + \|\mathrm{x}_\mathrm{q}^{(\mathrm{L})} - \mathrm{x}_\mathrm{q}^{(\ell)}\|_2 \|\mathrm{x}_\mathrm{c}^{(\mathrm{L})} - \mathrm{x}_\mathrm{c}^{(\ell)}\|_2$$
$$\triangleq \underbrace{\underbrace{\langle \mathrm{z}_\mathrm{q}^{(\ell)}, \mathrm{z}_\mathrm{c}^{(\ell)} \rangle}_{known} + \underbrace{\sqrt{(1-\|\mathrm{z}_\mathrm{q}^{(\ell)}\|_2^2)(1-\|\mathrm{z}_\mathrm{c}^{(\ell)}\|_2^2)}}_{known!\,\text{☺}}}_{U_{q,c}^{\ell}\ \ upper\text{-}bound\ at\ level-\ell} \quad (6)$$

3.1.2 Efficient Candidate Filtering using $U_{q,c}^{\ell}$

Using this upper bound, Pyramid Rank (Algorithm 1) proceeds as follows. We define an index-set $\mathcal{I}$ as the set of indices of all candidates in the database. Our goal is to filter this set and obtain top-K candidates. Given a query q, we first obtain embeddings of length d at level l for both the query q and all candidates c_i. We then compute the upper-bound U_{q,c_i}^{ℓ} for all candidates. Now, we filter out candidates with an upper-bound lower than threshold τ, thus leaving us with an updated index-set $\hat{I}$. The threshold τ is set as the mid-point of a range $(\tau_{\min}, \tau_{\max})$. Depending on the size of $\hat{I}$, we proceed to increase the Matryoshka representation level ℓ. If there are more than K elements in $\hat{I}$, it means we need more fine-grained representations (higher level ℓ) to distinguish between the candidates and rank them in order of decreasing relevance. In such a case, we tighten the threshold,

[1]Detailed derivation in Appendix.

Algorithm 1: Pyramid Rank - *Fast Candidate Filtering*

Input: Query q, database $\mathcal{C}=\{c_n\}_{n=1}^{N}$, target K, tolerance ϵ
Output: ϵ-tolerated top-K candidates in $\mathcal{C}$

Init: $\tau_{\min}=-1,\ \tau_{\max}=1,\ \ell=1,\ \mathcal{I}=\{1\ldots N\},\ \mathbf{U}_q=\text{zeros}(N)$
while $\tau_{\max}-\tau_{\min}>\epsilon$ **do**
  $\tau=(\tau_{\min}+\tau_{\max})/2$ // threshold
  Compute $\mathrm{z}_{\mathrm{q}}^{(\ell)}, \mathrm{z}_{\mathrm{c_i}}^{(\ell)}\ \forall i\in\mathcal{I}$ with q, c_i // using Eq. (2),Eq. (4)
  Compute $U_{q,c_i}^{\ell}\ \forall i\in\mathcal{I}$ with $\mathrm{z}_{\mathrm{q}}^{(\ell)}, \mathrm{z}_{\mathrm{c_i}}^{(\ell)}$ // using Eq. (6)
  $\hat{\mathcal{I}}=\{i\in\mathcal{I}\mid U_{q,c_i}^{\ell}\geq\tau\}$
  if $|\hat{\mathcal{I}}|\geq K$ **then**
    $\tau_{\min}=\tau$ // tighten threshold
    if $\ell<L$ **then**
      $\ell=\ell+1$
    $\mathcal{I}=\hat{\mathcal{I}}$ // update index set
    $\mathbf{U}_q[i]=U_{q,c_i}^{\ell}\ \forall\, i\in\mathcal{I}$ // update upper-bound vector
  else
    $\tau_{\max}=\tau$ // loosen threshold
return $\mathcal{R}_1=\{c_i\mid i\in(\text{argsort}^{\downarrow}\mathrm{U_q})[1{:}K]\}$ // return top-K

increase the representation level, and repeat the process; else we loosen the threshold and repeat the process. While iterating, we store the indices of the top-K candidates in index-set $\mathcal{I}$ and their corresponding similarity upper-bounds in vector U_q. Thus Pyramid Rank, through this multi-scale representation search, returns the ranked set of top-K candidates, sorted by their similarity upper-bound.

3.1.3 Implications of the Upper-Bound

Efficient Similarity Computation. Eq. (6) gives us an upper-bound on the similarity of the highest level-L representations (fine-grained, longer vectors), in terms of lower level-ℓ representations $\mathrm{z}_{\mathrm{c}}^{(\ell)}$ and $\mathrm{z}_{\mathrm{q}}^{(\ell)}$ (coarse, smaller vectors). For each query, q and c representations and similarity upper-bounds U_{q,c_i}^{ℓ} are computed online. These representations do not need any re-computation and are instead obtained by choosing the appropriate Matryoshka representation level and slicing the pre-computed full-length vector using Eq. (2). The Stage-1 algorithm allows us to use smaller vectors for candidates with low relevance, thus making the ranking process more efficient in comparison to naïvely using the full-length vectors for all candidates.

Computation Cost Derivation. Naïvely computing similarity between query and candidate at the highest MRL level would result in $\text{cost}=Nd(2^{L-1})$. However, with our algorithm, for a non-zero number of candidates, similarity computations are performed at level $\ell<L$, guaranteeing fewer computations than naïve MRL. For e.g., assuming a *uniform distribution* of levels $\ell\in\{1\ldots L\}$, we get $\text{cost}=\frac{Nd}{L}\sum_{\ell=1}^{L}2^{\ell-1}=\frac{Nd}{L}(2^{L-1}-1)$, i.e., L times fewer computations than naïve MRL. The probability distribution of levels chosen by Pyramid Rank depends on datasets, as shown in Fig. 11(b).

Admissibility Guarantee. Recall that a candidate is removed from index-set $\mathcal{I}$ if $U_{q,c}^{\ell}<\tau$, i.e., the upper-bound at level ℓ is less than the threshold. If a candidate is removed, Eq. (6) implies that the true similarity at level-L $\langle \mathrm{x}_{\mathrm{q}}^{(\mathrm{L})}, \mathrm{x}_{\mathrm{c}}^{(\mathrm{L})}\rangle \leq U_{q,c}^{\ell}<\tau$, i.e. no item's relevance is underestimated on account of the upper-bound.

Convergence Guarantee. At the beginning of the algorithm, the threshold interval is $w_0=\tau_{\max}-\tau_{\min}$. Each bisection step halves this width – to reach an ϵ-tolerance interval requires at most $\lceil\log_2(w_0/\epsilon)\rceil$ iterations, independent of the database size N, thus ensuring predictable runtime and stable termination across all database candidates and queries.

3.2 Stage-2: Multimodal Re-ranking

Stage-1 efficiently filters top-K candidates using text-only embeddings. The goal of Stage-2, therefore, is to refine the embeddings of the top-K retrieved items in $\mathcal{R}_1$ by re-introducing multimodal features to obtain fine-grained contextual representations for re-ranking. As Stage-2 operates without data modality conversion, it avoids information loss and exploits rich *cross-modal semantic similarity*, resulting in more accurate refinement of the top-K candidates without additional supervision.

Bidirectional-CoT Embedding Score. We design a prompting strategy to produce embeddings that denote shared alignment between queries and candidates. Given a query q and candidate $c_k\in\mathcal{R}_1$, we perform few-shot chain-of-thought (CoT) prompting to obtain a query-to-candidate alignment embedding z_{q2c} and a candidate-to-query alignment embedding z_{c2q} (Fig. 3). If all concepts in a query are found in a candidate, we denote this as a candidate with high alignment to the query and low alignment otherwise and vice versa. Using the above explanations in the CoT examples, we then pass queries and candidates in (i) the query-to-candidate (q2c) prompt (ii) the candidate-to-query (c2q) prompt. We add an `<emb>` token to the end of both prompts to compute embedding with a autoregressive generative model with explicit one-word limitation following (Liu et al., 2025; Jiang et al., 2024b,a). We extract the last hidden state preceding the `<emb>` token to use as embeddings (z_{q2c} and z_{c2q}), and compute the Bidirectional-CoT Embedding Score between z_{q2c} and z_{c2q} as:

$$S_{\text{CoT}}(q,c_k)=\text{COSINE}\big(z_{q2c}(q,c_k),z_{c2q}(c_k,q)\big) \quad (7)$$

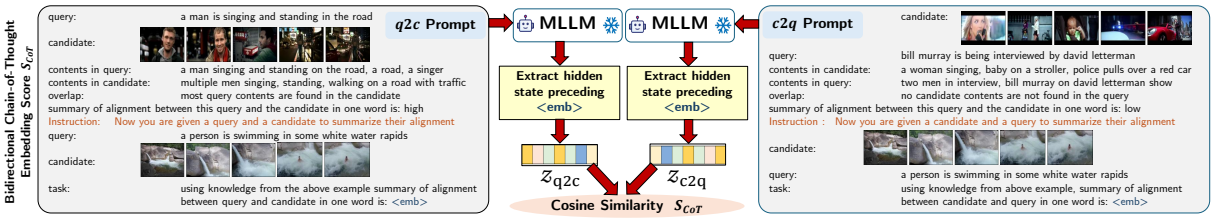

Figure 3: **Stage-2 Bidirectional-CoT Embedding Score** S_{CoT}: With CoT examples and a inserted *end-of-sequence* <emb> token; the hidden state preceding <emb> yields embeddings z_{q2c} and z_{c2q} for the q2c and c2q prompts, whose cosine similarity forms S_{CoT}. Unlike supervised methods, our bidirectional embeddings - one from query and another from candidate, are computed zero-shot for multimodal alignment with similarity computation.

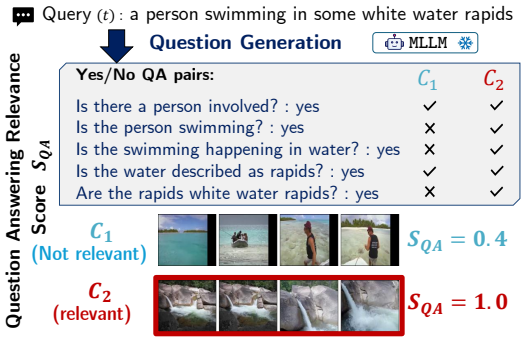

Figure 4: **Stage-2 Question Answering Relevance Score:** from the query, we generate a set of yes/no questions and compute QA accuracy using an MLLM on each database candidate to obtain S_{QA}.

Question Answering Relevance Score. Given a multimodal query q, we prompt an MLLM to generate M discriminative (yes/no) question-answer pairs $\{(\texttt{que}^m, \texttt{ans}^m)\}_{m=1}^{M}$, where each $\texttt{ans}^m \in \{\text{Yes}, \text{No}\}$.

Prompt for QA generation from Query:
```
You are given a multimodal query:      <q>
Generate M Yes/No questions that capture
essential semantics from the above query.
For each question, also provide its correct
Yes/No answer.  Format:
Q1:..?  A: Yes/No; Q2:..?  A: Yes/No; ..
```

For query q, we use an MLLM as a multimodal question answering model M to ask these questions with each top-K candidate as context. For each candidate, the MLLM yields a predicted answer $\text{MLLM}(\texttt{que}^m, c_k)$ which we compare with ground truth $\texttt{ans}^m$ to compute accuracy which we use as Question Answering Relevance Score:

$$S_{\text{QA}}(q, c_k) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\big(\texttt{ans}^m = \text{MLLM}(\texttt{que}^m, c_k)\big) \quad (8)$$

A candidate's QA logs provide explanation for *why* a candidate is retrieved or rejected.

Putting it together. For each top-K candidate c_k, we compute both scores independently and aggregate them through a convex combination with coefficient $\alpha \in [0, 1]$:

$$S_{\text{re-rank}}(q, c_k) = \alpha\, S_{\text{CoT}}(q, c_k) + (1 - \alpha)\, S_{\text{QA}}(q, c_k) \quad (9)$$

4 Experiments

4.1 Experimental Setup

We perform experiments on 12 types of MMIR tasks across 4 modalities (text t, image i, video v, audio a) and their combinations.

Datasets. For MMIR tasks involving images and text, we test on M-BEIR (Wei et al., 2024) (combining 10 retrieval tasks), Urban-1K (Zhang et al., 2024a), Flickr30K (Plummer et al., 2015), and GeneCIS (Vaze et al., 2023) benchmarks. For video-based retrieval ($t \to v$ and $v \to t$), we test on MSRVTT-1kA (Xu et al., 2016), MSVD (Chen and Dolan, 2011), LSMDC (Rohrbach et al., 2015), and DiDeMo (Hendricks et al., 2017). For audio retrieval ($t \to a$ and $t \to (a, v)$) we use AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020) benchmarks. These 32 datasets cover 12 distinct MMIR tasks and when combined have $5.7M$ candidates and $250K$ queries.

Evaluation Metrics. We use standard evaluation metric Recall at k (R@k), which denotes % of queries with the relevant item in top-K. Following prior work (Oncescu et al., 2021; Reddy et al., 2025; Liu et al., 2025; Kim et al., 2025) we evaluate R@1, R@5, R@10 and nDCG@10[2].

Implementation Details. For Stage-1, we use Qwen3-MRL (Zhang et al., 2025b) as the feature extractor to obtain multi-scale Matryoshka representations. For Stage-2 re-ranking, the Bidirectional-CoT Embedding Score and Question Answering Relevance Score are computed using Qwen2.5-Omni-7B without any fine-tuning to ensure a completely zero-shot evaluation. For image retrieval tasks we use α=0.6 while for video and audio retrieval tasks we use α=0.3 in Eq. (9).

	Methods	$t \to i$			$t \to t$	$t \to (i,t)$		$i \to t$			$i \to i$	$(i,t) \to t$		$(i,t) \to i$		$(i,t) \to (i,t)$	
		VisualNews R@5	MSCOCO R@5	Fashion200K R@10	WebQA R@5	EDIS R@5	WebQA R@5	VisualNews R@5	MSCOCO R@5	Fashion200K R@10	NIGHTS R@5	OVEN R@5	InfoSeek R@5	FashionIQ R@10	CIRR R@5	OVEN R@5	InfoSeek R@5
Zero-Shot	CLIP-L (Radford et al., 2021)	43.3	61.1	6.6	36.2	43.3	45.1	41.3	79.0	7.7	26.1	24.2	20.5	7.0	13.2	38.8	26.4
Zero-Shot	SigLIP (Zhai et al., 2023)	30.1	75.7	36.5	39.8	27.0	43.5	30.8	88.2	34.2	28.9	29.7	25.1	14.4	22.7	41.7	27.4
Zero-Shot	Qwen3-Omni-30B-Th (Xu et al., 2025b)	16.6	59.9	12.5	46.9	31.7	14.6	10.9	50.5	9.7	28.6	26.7	25.7	7.5	17.0	49.2	37.9
Zero-Shot	Qwen3-MRL (Kusupati et al., 2022)	32.5	57.2	19.0	50.3	39.8	47.1	20.4	55.0	15.7	32.7	28.5	26.3	25.4	21.2	54.0	42.8
Zero-Shot	MULTI-SCORE	**49.8**	**86.9**	**41.8**	**77.4**	**79.3**	**81.1**	**49.7**	**92.3**	**38.4**	**36.0**	**50.4**	**62.7**	**38.5**	**58.3**	**68.7**	**54.9**
Supervised	UniR-BLIP$_{FF}$ (Wei et al., 2024)	23.4	79.7	26.1	80.0	50.9	79.8	22.8	89.9	28.9	33.0	41.0	22.4	29.2	52.2	55.8	33.0
Supervised	UniR-CLIP$_{SF}$ (Wei et al., 2024)	42.6	81.1	18.0	84.7	59.4	78.7	43.1	**92.3**	18.3	32.0	45.5	27.9	24.4	44.6	67.6	48.9
Supervised	GENIUS$^{\mathcal{R}}$ (Kim et al., 2025)	27.4	78.0	16.2	44.6	44.3	60.6	28.4	91.1	16.3	30.2	41.9	20.7	19.3	39.5	52.5	30.1
Supervised	LamRA-ret (Liu et al., 2025)	41.6	81.5	28.7	86.0	62.6	81.2	39.6	90.6	30.4	32.1	54.1	52.1	33.2	53.1	76.2	63.3
Supervised	LamRA (Liu et al., 2025)	**48.0**	**85.2**	**32.9**	**96.7**	**75.8**	**87.7**	**48.6**	**92.3**	**36.1**	**33.5**	**59.2**	**64.1**	**37.8**	**63.3**	**79.2**	**78.3**

Table 1: Results on the M-BEIR benchmark (Wei et al., 2024) with each dataset and its standard evaluation metric. Colors: best zero-shot results: **bold blue**; best supervised results: **bold red**; best overall results: highlighted yellow.

	Methods	$t \to i$		$i \to t$		$(i,t) \to i$
		Urban-1K	Flickr30K	Urban-1K	Flickr30K	GeneCIS
Zero-Shot	CLIP-L (Radford et al., 2021)	52.8	67.3	68.7	87.2	13.3
Zero-Shot	Long-CLIP-L (Zhang et al., 2024a)	86.1	76.1	82.7	89.3	16.3
Zero-Shot	UniIR-CLIP (Wei et al., 2024)	75.0	78.7	78.4	94.2	16.8
Zero-Shot	E5-V (Jiang et al., 2024b)	84.0	79.5	82.4	88.2	18.5
Zero-Shot	MagicLens-L (Zhang et al., 2024b)	59.3	72.5	24.2	84.6	16.3
Zero-Shot	EVA-CLIP-18B (Sun et al., 2024)	81.7	83.3	83.3	96.7	13.6
Zero-Shot	Qwen3-Omni-30B-Th (Xu et al., 2025b)	56.9	62.0	54.6	68.0	14.6
Zero-Shot	Qwen3-MRL (Kusupati et al., 2022)	67.3	69.4	63.6	73.9	19.5
Zero-Shot	MULTI-SCORE	**98.2**	**89.6**	**97.8**	**98.1**	**34.6**
Sup.	LamRA-ret (Liu et al., 2025)	95.1	82.8	94.3	92.7	18.9
Sup.	LamRA (Liu et al., 2025)	**98.8**	**88.1**	**98.0**	**97.6**	**24.8**

Table 2: Additional results on image-text retrieval tasks (R@1) on Urban-1K, Flickr30K, and GeneCIS.

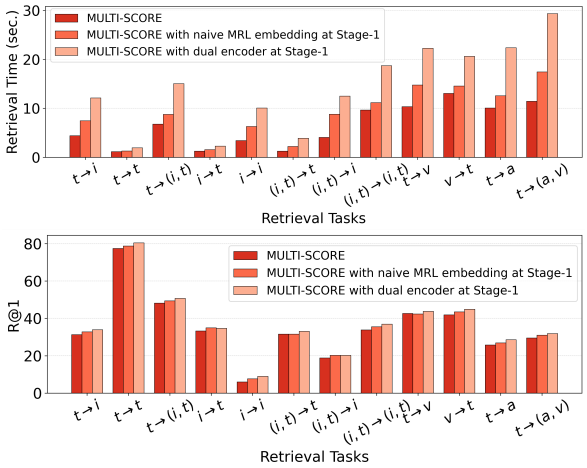

Figure 5: Retrieval performance and inference time using (1) MULTI-SCORE, (2) MULTI-SCORE with naïve full-resolution MRL embedding at Stage-1, and (3) MULTI-SCORE with computationally expensive dual encoder at Stage-1. Expensive rankers in Stage-1 can marginally improve performance but at an impractical inference cost - hurting adaptability.

Baseline Models[2]**.** Although MULTI-SCORE is zero-shot and training-free, we compare with both zero-shot and supervised methods. For image retrieval, zero-shot baselines include widely used dual encoder models and VLMs such as CLIP and Qwen-family models[2], while supervised baselines include LamRA (Liu et al., 2025), Genius$^{\mathcal{R}}$ (Kim et al., 2025), UniR (Wei et al., 2024) with different encoders (CLIP, BLIP) and fusion strategies (score-level (SF), feature-level (FF)). For zero-shot video and audio retrieval, we compare with Qwen3-Omni-30B-Thinking (Th) (Xu et al., 2025b), LamRA, Qwen2-Audio-7B (Chu et al., 2024); and supervised MCQ-BridgeFormer (Ge et al., 2022), OA-Trans (Wang et al., 2022), NeighborRetr (Lin et al., 2025), SOTA (Oncescu et al., 2021), AVR (Nagrani et al., 2022).

4.2 Results

In task-specific information retrieval, performance is evaluated on a single dataset, where the candidate pool consists solely of items from that dataset.

Results on text-image benchmarks. In Tab. 1, MULTI-SCORE is compared against dual encoder models CLIP (Radford et al., 2021), Long-CLIP (Zhang et al., 2024a), BLIP (Li et al., 2022), SigLIP (Zhai et al., 2023) for zero-shot evaluation. MULTI-SCORE outperforms existing zero-shot MMIR systems on M-BEIR benchmark on 8 different retrieval tasks, with notable improvements over existing zero-shot baselines: improvement by 37% on InfoSeek, 36% on WebQA. Although MULTI-SCORE is zero-shot, it still outperforms state-of-the-art supervised methods on several datasets: VisualNews, MSCOCO, EDIS, NIGHTS, FashionIQ. In Tab. 2 we show performance on 5 more text-image retrieval datasets, spanning 3 different tasks, and observe similar trends; large gains w.r.t. zero-shot models and outperforming state-of-the-art supervised method LamRA in FLickr30K and GenCIS.

Results on video benchmarks. In Tab. 3 (left), we compare video retrieval performance with zero-shot methods which are fine-tuned on auxiliary video understanding tasks (e.g., MCQ-BridgeFormer, OA-Trans, and also with MLLMs such Qwen3-Omni. For video-based retrieval, MULTI-SCORE outperforms zero-shot, supervised methods, and naïve Qwen3-MRL on all datasets, for instance, 45 points supervised performance gain

[2]Additional benchmarks, baselines, metrics in Appendix.

	Methods	$t \rightarrow v$ MSRVTT-1kA	MSVD	LSMDC	DiDeMo	$v \rightarrow t$ MSRVTT-1kA	MSVD	LSMDC	DiDeMo
Zero-Shot	MCQ-BridgeFormer (Ge et al., 2022)	26.0	43.6	12.2	25.6	–	–	–	–
	OA-Trans (Wang et al., 2022)	23.4	–	–	23.5	–	–	–	–
	LamRA (Liu et al., 2025)	44.7	52.4	–	–	–	–	–	–
	Qwen3-Omni-30B-Th (Xu et al., 2025b)	39.5	51.0	21.1	37.3	48.5	26.8	62.2	34.1
	Qwen3-MRL (Kusupati et al., 2022)	33.3	37.6	15.6	31.9	30.5	21.2	39.7	28.5
	MULTI-SCORE	55.7	69.4	27.1	50.3	53.1	31.9	68.0	50.7
Supervised	MCQ-BridgeFormer (Ge et al., 2022)	37.6	52.0	17.9	37.0	–	–	–	–
	OA-Trans (Wang et al., 2022)	35.8	39.1	18.2	34.8	17.5	–	–	–
	T2VLAD (Wang et al., 2021)	29.5	–	14.3	–	31.8	–	14.2	–
	X-Pool (Gorti et al., 2022)	46.9	47.2	25.2	–	43.9	66.4	22.7	–
	RPC (Lai et al., 2025)	47.3	38.5	22.8	34.7	46.3	48.1	22.0	33.4
	NeighborRetr (Lin et al., 2025)	49.5	47.9	–	48.2	48.7	63.3	–	48.4

	Methods	$t \rightarrow a$ AudioCaps	Clotho	$t \rightarrow (a, v)$ AudioCaps
Zero-Shot	AVR-VC3M (Nagrani et al., 2022)	8.7	10.6	3.0
	Qwen2-Audio-7B (Chu et al., 2024)	29.7	18.3	23.5
	Qwen2.5-Omni-7B (Xu et al., 2025a)	31.1	18.5	30.9
	Qwen3-Omni-30B-Th (Xu et al., 2025b)	32.1	15.4	21.8
	Qwen3-MRL (Kusupati et al., 2022)	35.4	17.9	22.7
	MULTI-SCORE	45.4	28.2	44.5
Sup.	SOTA (Oncescu et al., 2021)	24.3	6.7	28.1
	AVR (Nagrani et al., 2022)	32.0	7.8	41.4
	AVR-VC3M (Nagrani et al., 2022)	35.5	8.4	43.2

Table 3: Performance on cross-modal text-video *(left)* and text-audio *(right)* retrieval in terms of R@1.

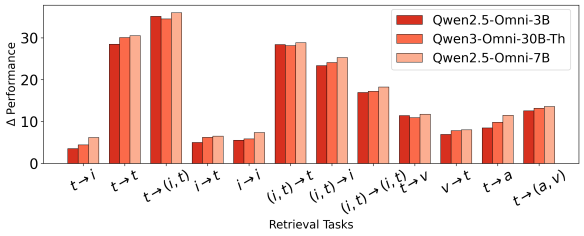

Figure 6: Backbone-agnostic performance gain (Δ) with MULTI-SCORE compared to existing SOTA zero-shot methods on all 12 tasks, using 3 MLLMs as Stage-2 backbones.

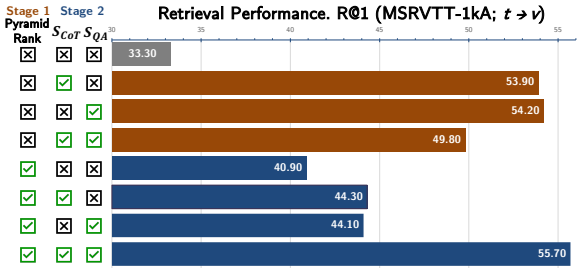

Figure 7: Each component of our method (Stages-1 and 2) improves the performance vs. using the full-scale (D=1024) Qwen3-MRL embeddings (top gray bar).

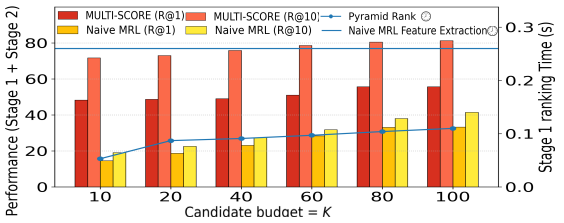

Figure 8: Effect of candidate budget (K) on retrieval performance and runtime. As K increases, re-ranking more candidates in Stage-2 improves retrieval performance, while at Stage-1 Pyramid Rank keeps computational cost low; achieving efficient retrieval compared to naïve MRL embeddings. Bars: recall; blue lines: time.

in LSMDC.

Results on audio benchmarks. Tab. 3 (right) shows that, MULTI-SCORE achieves the highest zero-shot performance across both AudioCaps and Clotho benchmarks, surpassing all full-resolution naïve Qwen3-MRL and Omni variants. MULTI-SCORE maintains strong performance in the joint audio-visual setting on AudioCaps, demonstrating consistent cross-modal retrieval strength.

Universal information retrieval. We evaluate retrieval accuracy and efficiency by constructing a unified candidate pool across all 32 datasets, requiring retrieval in a universal cross-modality search space. Fig. 5 shows that instead of using our proposed Pyramid Rank, computing full-scale naïve MRL embedding at Stage-1 only improves performance marginally, but at a significant inference cost. Using dual encoders CLIP for images, CLIP4Clip (Luo et al., 2022) for videos, AudioCLIP (Guzhov et al., 2022) for audio in Stage-1 improves performance, though at even higher inference cost. Using Pyramid Rank in Stage-1 therefore yields optimal performance, and efficiency.

4.3 Analysis

Backbone Agnostic Performance Gain. Fig. 6 indicates that Qwen2-Omni-7B shows the most significant zero-shot performance gain in comparison to other backbones for all tasks. However, regardless of the choice of backbone model in Bidirectional-CoT Embedding Score and Question Answering Relevance Score, Stage-2 re-ranking always improves performance.

Ablation Study. Fig. 7 analyzes the contribution of each component in MULTI-SCORE on the MSRVTT-1kA benchmark. Combining all three modules: Pyramid Rank, S_{CoT}, and S_{QA} results in the highest retrieval performance (R@1 = 55.7), demonstrating their complementary impact.

Efficiency of Pyramid Rank. In Stage-1, Pyramid Rank retrieves top-K candidates that Stage-2 uses for re-ranking. The runtime of Stage-1 increases with K. However, the bigger the search space is, the more effective Stage-2 re-ranking will be. Fig. 8 confirms this: as we re-rank more candidates in Stage-2, the retrieval performance improves and re-ranking more candidates in Stage-2 is computationally cheap when we use Pyramid Rank in Stage-1 for ranking. Thereby, Pyramid Rank im-

Caption Len.	Image R@1	Video R@1	Audio R@1
(0 , 300]	52.1	47.4	38.5
(300 , 600]	54.3	49.8	39.1
> 600	**61.8**	**52.7**	**39.8**

(a) Longer captions (> 600 tokens) help retrieve across all modalities.

# Questions	Image R@1	Video R@1	Audio R@1
(0 , 3]	51.9	46.3	37.2
(3 , 6]	56.8	50.4	38.9
≥ 7	**58.4**	**51.1**	**42.8**

(b) Using more questions help retrieve across all modalities.

# CoT examples	Image R@1	Video R@1	Audio R@1
2	**57.5**	**50.8**	**39.4**
3	57.3	50.5	39.0
4	57.4	50.2	39.1

(c) Using 2 CoT examples achieves optimal retrieval across modalities.

Table 4: Effect of caption length, # questions, and # CoT examples on retrieval performance: R@1 (↑).

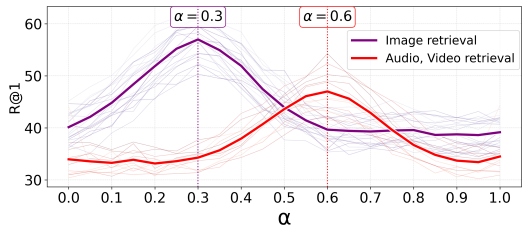

Figure 9: The faded lines represent R@1 for different values of α on all 32 retrieval datasets, while the bold lines indicate R@1 averaged across datasets. Optimal zero-shot performance is observed at $\alpha = 0.3$ for image retrieval and $\alpha = 0.6$ for audio/video retrieval.

Pyramid Rank Backbone	Captioning?	R@100	Hit@100	Inference Time (sec)
ResNet-MRL	No	73.3	94.4	0.065
Qwen3-MRL	Yes	73.0	92.1	0.072

Table 5: Pyramid Rank is robust to text conversion, preserving R@100, Hit@100, and runtime.

Feature Extraction Model	Feature Storage (KB)
I3D (Carreira and Zisserman, 2017)	720
X3D-XL (Feichtenhofer, 2020)	980
Two-Stream I3D (Carreira and Zisserman, 2017)	1440
Two-Stream S3D-G (Xie et al., 2018)	850
SlowFast 16×8 + R101 + NL (Feichtenhofer et al., 2019)	1360
MULTI-SCORE Captioning (QWEN3-VL-8B) **(Ours)**	**16.7**

Table 6: Captioning based pre-processing helps efficient offline feature storage for video database.

proves efficiency by filtering candidates faster for Stage-2 re-ranking. Using naïve MRL embeddings without re-ranking yields sub-optimal performance, illustrating the importance of Stage-2.

Hyperparameter Analysis. Fig. 9 shows the R@1 on 8 image retrieval tasks and 4 audio and video retrieval tasks for different α values in Eq. (9). $\alpha = 0.3$ results in peak zero-shot performance in image retrieval and $\alpha = 0.6$ for audio and video retrieval. Tab. 4 shows three factors that aid that Bidirectional-CoT Embedding Score help image, video, and audio retrieval performance: longer captions as input for Pyramid Rank, more questions in Question Answering Relevance Score, and two CoT examples in Fig. 11 (a) shows as tolerance ϵ in Algorithm 1 decreases, both retrieval performance and inference time decreases (i.e., Stage-2 receives fewer and less relevant candidates, as the search becomes overly conservative and eliminates potential matches too early). Results in Tab. 4, Figs. 9 and 11 (a) are averaged across all 32 datasets.

Visualizing Dataset Difficulty. Fig. 11 (b) visualizes the distribution of upper bounds estimated by Pyramid Rank for datasets of varying difficulty: FashionIQ (difficult), WebQA (medium), and MSCOCO (easy). Easier datasets have tighter and more right-shifted upper bounds, indicating stronger similarity scores between queries and relevant candidates, while harder datasets like FashionIQ show broader and left-skewed distributions, reflecting more ambiguous candidate similarity.

Qualitative Results. Fig. 10 compares qualitative retrieval outcomes across MULTI-SCORE (ours), its naïve MRL variant, and the supervised LamRA baseline for two different multimodal queries. Our two-stage framework efficiently retrieves targets in 0.11 s with higher semantic accuracy than competing methods, thus demonstrating its fine-grained, efficient, zero-shot retrieval capability. While the naïve MRL variant retrieves visually similar yet less precise candidates with higher latency (0.31 s), LamRA attains moderate accuracy but at the expense of domain-specific training.

Captioning based pre-processing has minimal impact on Retrieval and helps efficient offline feature storage. Pyramid Rank is modality-agnostic, but sicce Qwen3-MRL can process only text (and is the only public foundation model with MRL), our design choice of converting queries and candidate to text (for e.g. via captioning for images and videos) reflects current model availability rather than an algorithmic limitation. In Tab. 5, using vision MRL features directly with ResNet-MRL versus using captioned Qwen3-MRL yields only a marginal drop in R@100 (73.26 vs. 72.98), while keeping Hit@100 high (94.41 vs. 92.07) and inference time comparable, showing that modality-to-text conversion does not substantially harm top-K candidate survival. Tab. 6 additionally demonstrates higher efficiency in storage than standard video feature pipelines: at 15 FPS on MSRVTT-1K, classic methods store 700-1400 KB per video, whereas MULTI-SCORE stores only 16.7 KB.

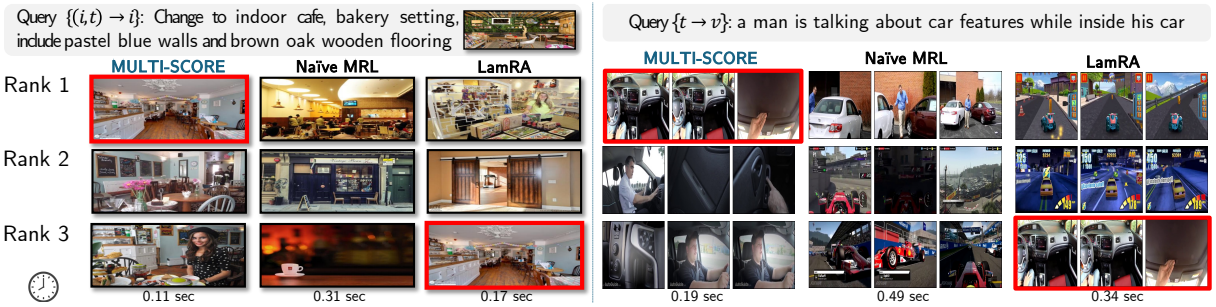

Figure 10: Qualitative Results: MULTI-SCORE consistently retrieves semantically aligned target (in red box) faster and more accurately than the Naïve MRL (only full scale MRL embedding based ranking) and LamRA (supervised) baselines. Unlike LamRA, which misinterprets scene context (e.g. retrieving images with different wall color), and Naïve MRL that retrieves videos with man outside a car, MULTI-SCORE effectively grounds textual edits within the visual context, highlighting its ability to perform precise and context-aware multimodal retrieval, with no training.

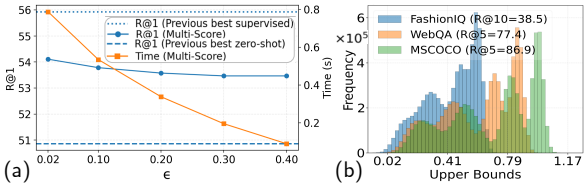

Figure 11: (a) Larger tolerance ϵ reduces Stage-1 inference time but slightly lowers R@1. At optimal $\epsilon = 0.02$, MULTI-SCORE catches up to previous best supervised performance while at the worst $\epsilon = 0.40$, MULTI-SCORE is still equivalent to previous best zero-shot methods. (b) Distribution of upper bounds estimated by the Pyramid Rank across datasets of varying difficulty: FashionIQ (hard), WebQA (medium), and MSCOCO (easy).

Captioning Model	Size	Prompt Type	R@1	R@5	R@10
BLIP-2	2.7B	Concise ("Describe the image")	80.9	85.1	89.4
		Detailed (Sec. 6.4)	81.7	85.8	89.6
Qwen3-VL	8B	Concise ("Describe the image.")	82.1	85.7	90.3
		Detailed (Sec. 6.4)	82.7	86.9	91.5

Table 7: Backbone model family, size, prompt variance have minimal impact on Stage-1 ranking performance.

Multi-Score performance gains are agnostic to backbone model family, size, and prompt sensitivity as demonstrated by results in Tabs. 7 to 9. These gains are driven by our proposed Pyramid Rank, Bidirectional-CoT Embedding Score and Question Answering Relevance Score, not from using bigger foundation models.

Hardware, Decoding, and Compute costs: from offline pre-processing to online inference. For Stage-1, we use 8 clusters of 4 NVIDIA H100 GPUs (batch size of 128), and for Stage-2, we use 4 clusters of 16 NVIDIA L40S GPUs (Bidirectional-CoT batch size of 8, QA scoring batch size of 4). The decoding setting is: temperature 0, QA Relevance scoring with 5 max tokens, CoT generation with <emb> token for hidden-state extraction;

Re-ranking Model	Size	Prompt Type	R@1	R@5	R@10
LLaVA-1.5	3B	without alignment in Fig. 3	76.4	84.9	90.2
		with alignment in Fig. 3 (ours)	80.3	86.2	91.0
Qwen2.5-Omni	7B	without alignment in Fig. 3	81.2	85.8	90.7
		with alignment in Fig. 3 (ours)	82.7	86.9	91.5

Table 8: Backbone model family, size, prompt variance minimally impact on Stage-2 re-ranking performance.

Stage-1 Backbone	Stage-2 Backbone	R@1
ResNet-MRL	None (embedding only)	64.5
ResNet-MRL	LLaVA-7B (standard MLLM re-ranking)	66.2
ResNet-MRL	MULTI-SCORE (LLaVA-7B CoT + QA)	**71.8**
Qwen3-VL-8B (MRL)	None (embedding only)	74.3
Qwen3-VL-8B (MRL)	LLaVA-7B (standard MLLM re-ranking)	77.1
Qwen3-VL-8B (MRL)	MULTI-SCORE (LLaVA-7B CoT + QA)	**82.7**

Table 9: MULTI-SCORE consistently improves over both weak and strong backbones, outperforming backbone-only retrieval and standard MLLM re-ranking.

CoT and QA are executed in parallel. On a 5.7M database, our total offline captioning cost is 20 GPU-hours (parallelized across 32 GPUs) and per-query online inference cost is 0.87 s[3].

5 Conclusion

We introduced MULTI-SCORE, a fine-tuning-free two-stage multimodal retrieval framework combining efficient hierarchical filtering with fine-grained multimodal alignment. By proposing Pyramid Rank, we are the first to leverage Matryoshka representations for pyramidal embedding-based filtering to achieve efficient multimodal retrieval. Through Bidirectional-CoT Embedding Score and Question Answering Relevance Score re-ranking with multimodal alignment, MULTI-SCORE results in state-of-the-art zero-shot performance across 12 MMIR tasks on 32 datasets, demonstrating its efficacy across tasks, domains, and modalities.

[3]Detailed breakdown of compute costs is in the Appendix

Limitations

While Pyramid Rank is inherently modality-agnostic and can operate with any foundation embedding model that supports a pyramidal representation hierarchy (for e.g. Matryoshka representations (Kusupati et al., 2022)), at the time of writing this paper, the only foundation embedding model which supports MRL embeddings is Qwen3-MRL. Qwen3-MRL operates on text-only inputs and therefore requires us to convert image, video, and audio data into text via captioning for use with our proposed Pyramid Rank approach. Importantly, this limitation arises from the current ecosystem of embedding models rather than from the proposed Pyramid Rank algorithm itself. This dependency may introduce information loss due to modality conversion and additional offline pre-processing (captioning) cost. Pyramid Rank algorithm can be seamlessly applied to future embedding models designed to produce pyramidal representations.

Acknowledgments

This work was funded in part by the Defense Advanced Research Projects Agency's (DARPA) SciFy program under agreement number HR00112520301. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views, opinions, and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of employers, funding agencies, or governments. We acknowledge high performance computing support from UMBC HPCF and a Lambda Inc. award to SS. We thank Reno Kriz for initial discussions on training-free retrieval and Frank Ferraro for feedback on the manuscript.

References

Edward H Adelson, Charles H Anderson, James R Bergen, Peter J Burt, and Joan M Ogden. 1984. Pyramid methods in image processing. *RCA engineer*, 29(6):33–41.

Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. 2021. Noise estimation using density estimation for self-supervised multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6644–6652.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738.

Harry G Barrow and Jay M Tenenbaum. 1981. Computational vision. *Proceedings of the IEEE*, 69(5):572–595.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. 2021. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Christoph Feichtenhofer. 2020. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *Advances in Neural Information Processing Systems*, 36:50742–50768.

Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. 2022. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16167–16176.

Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. 2022. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5006–5015.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE.

Efstathios Hadjidemetriou, Michael D Grossberg, and Shree K Nayar. 2004. Multiresolution histograms and their use for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7):831–847.

Godfrey Harold Hardy, John Edensor Littlewood, and George Pólya. 1952. *Inequalities*. Cambridge university press.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2024a. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196.

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

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of NAACL-HLT*, pages 119–132.

Sungyeon Kim, Xinliang Zhu, Xiaofan Lin, Muhammet Bastan, Douglas Gray, and Suha Kwak. 2025. Genius: A generative framework for universal multimodal search. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19659–19669.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and 1 others. 2022. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249.

Huakai Lai, Guoxin Xiong, Huayu Mai, Xiang Liu, and Tianzhu Zhang. 2025. Rethinking noisy video-text retrieval via relation-aware alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9231–9241.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 2169–2178. IEEE.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

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.

Yongqi Li, Wenjie Wang, Leigang Qu, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. 2024. Generative cross-modal retrieval: Memorizing images in multimodal language models for retrieval and beyond. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11851–11861.

Zengrong Lin, Zheng Wang, Tianwen Qian, Pan Mu, Sixian Chan, and Cong Bai. 2025. Neighborretr: Balancing hub centrality in cross-modal retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9263–9273.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. 2025. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4015–4025.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.

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.

David Marr. 1982. Vision: A computational investigation into the human representation and processing of visual information.

Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and

Cordelia Schmid. 2022. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pages 407–426. Springer.

Andreea-Maria Oncescu, A Koepke, João F Henriques, Zeynep Akata, and Samuel Albanie. 2021. Audio retrieval with natural language queries. In *Proc. Interspeech 2021*, pages 2411–2415.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

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*, pages 8748–8763. PmLR.

Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M de Melo, Benjamin Van Durme, and Rama Chellappa. 2025. Video-colbert: Contextualized late interaction for text-to-video retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19691–19701.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212.

Hermann Amandus Schwarz. 1890. Ueber ein die flächen kleinsten flächeninhalts betreffendes problem der variationsrechnung: Festschrift zum siebzigsten geburtstage des herrn karl weierstrass. In *Gesammelte Mathematische Abhandlungen: Erster Band*, pages 223–269. Springer.

Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. 2022. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 20020–20029.

Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. 2024. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*.

Qwen Team. 2025a. Qwen2.5-vl.

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

Sagar Vaze, Nicolas Carion, and Ishan Misra. 2023. Genecis: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6862–6872.

Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2022. Object-aware video-language pre-training for retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3313–3322.

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

Xiaohan Wang, Linchao Zhu, and Yi Yang. 2021. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5079–5088.

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

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, and Xinfa Zhu. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging

video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Yang Yuan. 2023. On the power of foundation models. In *International conference on machine learning*, pages 40519–40530. PMLR.

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.

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.

Biao Zhang, Lixin Chen, Tong Liu, and Bo Zheng. 2025a. Smec: Rethinking matryoshka representation learning for retrieval embedding compression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26220–26233.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. 2024b. Magiclens: Self-supervised image retrieval with open-ended instructions. In *International Conference on Machine Learning*, pages 59403–59420. PMLR.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yidan Zhang, Ting Zhang, Dong Chen, Yujing Wang, Qi Chen, Xing Xie, Hao Sun, Weiwei Deng, Qi Zhang, Fan Yang, and 1 others. 2024c. Irgen: Generative modeling for image retrieval. In *European Conference on Computer Vision*, pages 21–41. Springer.

Appendix

MULTI-SCORE is an abbreviation of the method name, but it also represents several innovations: multi-scale similarity computation (Stage-1), multimodal contextual representations (Stage-2), and multiple scoring functions for re-ranking (Stage-2).

6 Pyramid Rank: Derivation and Analysis

We design Pyramid Rank (Algorithm 1) to ensure the following benefits:

- **Admissibility:** no relevant candidate is ever pruned because the upper bound is admissible.
- **Convergence:** guaranteed termination in logarithmic steps, independent of database size.
- **ϵ-bounded correctness:** performance drop in comparison to naïvely computing query-candidate similarity using highest MRL level for the entire database is at most ϵ, which is controllable.

6.1 Detailed Derivation of Upper-Bound for Similarity

We use Qwen3-MRL 0.6B Embedding model (Zhang et al., 2025b) to compute embeddings x_c for all database entries c and x_q for input queries and normalize them to be unit-norm. This MRL (Matryoshka Representation Learning) feature extractor produces embeddings at L=6 levels, with the smallest length d=32 (at level $\ell = 1$) and largest length D=1024 (at level L=6) with the following properties:

$$\mathrm{x}_{\mathrm{c}}^{(\ell)} \in \mathbb{R}^{2^{\ell-1}d} \text{ ... length of level-}\ell \text{ vectors} \quad (10)$$

$$\mathrm{x}_{\mathrm{c}}^{(\ell)} = \mathrm{x}_{\mathrm{c}}^{(\mathrm{L})}[1 : 2^{\ell-1}d] \text{ ... nested MRL embeddings} \quad (11)$$

$$\|\mathrm{x}_{\mathrm{c}}^{(\mathrm{L})}\|_2 = \|\mathrm{x}_{\mathrm{q}}^{(\mathrm{L})}\|_2 = 1 \text{ ... unit-norm level-}L \text{ vectors} \quad (12)$$

We create zero-padded candidate and query embeddings $\mathrm{z}_{\mathrm{c}}^{(\ell)}, \mathrm{z}_{\mathrm{q}}^{(\ell)} \in \mathbb{R}^D$ as:

$$\mathrm{z}_{\mathrm{c}}^{(\ell)} = \mathrm{concat}(\mathrm{x}_{\mathrm{c}}^{(\ell)}, zeros(D - 2^{\ell-1}d)); \text{and}$$
$$\mathrm{z}_{\mathrm{q}}^{(\ell)} = \mathrm{concat}(\mathrm{x}_{\mathrm{q}}^{(\ell)}, zeros(D - 2^{\ell-1}d)). \quad (13)$$

For retrieval, we ideally want to compute similarity of each candidate vector $\mathrm{x}_{\mathrm{c}}^{(\mathrm{L})}$ with the query vector $\mathrm{x}_{\mathrm{q}}^{(\mathrm{L})}$, at the highest (most fine-grained) representation level L. However, this is computationally expensive, with a complexity of $O(2^L N)$. Therefore, Stage-1 seeks to reduce this computational complexity by leveraging multi-scale MRL representations. We derive an upper bound for $\langle \mathrm{x}_{\mathrm{q}}^{(\mathrm{L})}, \mathrm{x}_{\mathrm{c}}^{(\mathrm{L})} \rangle$ in

terms of lower-dimensional $x_c^{(\ell)}, x_q^{(\ell)}$, thus significantly reducing the cost associated with similarity computation.

$$\langle x_q^{(L)}, x_c^{(L)} \rangle = \underbrace{\langle z_q^{(\ell)}, z_c^{(\ell)} \rangle}_{known} + \underbrace{\langle x_q^{(L)} - z_q^{(\ell)}, x_c^{(L)} - z_c^{(\ell)} \rangle}_{unknown\text{☹}} \quad (14)$$

Applying the Cauchy-Schwartz inequality (Hardy et al., 1952) on the *unknown* term:

$$\langle x_q^{(L)}, x_c^{(L)} \rangle \leq \langle z_q^{(\ell)}, z_c^{(\ell)} \rangle + \|x_q^{(L)} - x_q^{(\ell)}\|_2 \|x_c^{(L)} - x_c^{(\ell)}\|_2 \quad (15)$$

From the definition of ℓ_2 vector norm:

$$\|x^{(L)}\|_2^2 = \langle x^{(L)}, x^{(L)} \rangle \quad (16)$$

Using Eq. (12) and Eq. (13), we can rewrite Eq. (16) as:

$$\|x^{(L)}\|_2^2 = \langle z^{(\ell)} + (x^{(L)} - z^{(\ell)}), z^{(\ell)} + (x^{(L)} - z^{(\ell)}) \rangle \quad (17)$$

Using the distributive property of vector inner product, we can rewrite Eq. (17) as:

$$\begin{aligned} \|x^{(L)}\|_2^2 = &\langle z^{(\ell)}, z^{(\ell)} \rangle + \langle (x^{(L)} - z^{(\ell)}), (x^{(L)} - z^{(\ell)}) \rangle + \\ &\langle z^{(\ell)}, (x^{(L)} - z^{(\ell)}) \rangle + \langle (x^{(L)} - z^{(\ell)}), z^{(\ell)} \rangle \end{aligned} \quad (18)$$

Using the symmetry of vector inner product, we can rewrite Eq. (18) as:

$$\begin{aligned} \|x^{(L)}\|_2^2 = &\langle z^{(\ell)}, z^{(\ell)} \rangle + \langle (x^{(L)} - z^{(\ell)}), (x^{(L)} - z^{(\ell)}) \rangle + \\ &2\langle z^{(\ell)}, (x^{(L)} - z^{(\ell)}) \rangle \end{aligned} \quad (19)$$

Since $z^{(\ell)}$ and $(x^{(L)} - z^{(\ell)})$ are orthogonal vectors, their dot product $\langle z^{(\ell)}, (x^{(L)} - z^{(\ell)}) \rangle = 0$, we can rewrite Eq. (19) as:

$$\|x^{(L)}\|_2^2 = \langle z^{(\ell)}, z^{(\ell)} \rangle + \langle (x^{(L)} - z^{(\ell)}), (x^{(L)} - z^{(\ell)}) \rangle \quad (20)$$

From Eq. (12), $\|x^{(L)}\|_2 = 1$, we can rewrite Eq. (20) as:

$$\langle z^{(\ell)}, z^{(\ell)} \rangle + \langle (x^{(L)} - z^{(\ell)}), (x^{(L)} - z^{(\ell)}) \rangle = 1 \quad (21)$$

From definition of ℓ_2 vector norm, we can rewrite Eq. (21) as:

$$\|x^{(L)} - z^{(\ell)}\|_2 = \sqrt{1 - \|z^{(\ell)}\|_2^2} \quad (22)$$

Finally, by plugging the value of $\|x^{(L)} - z^{(\ell)}\|_2$ into Eq. (15), we get:

$$\begin{aligned} \langle x_q^{(L)}, x_c^{(L)} \rangle &\leq \langle z_q^{(\ell)}, z_c^{(\ell)} \rangle + \|x_q^{(L)} - x_q^{(\ell)}\|_2 \|x_c^{(L)} - x_c^{(\ell)}\|_2 \\ &\triangleq \underbrace{\langle z_q^{(\ell)}, z_c^{(\ell)} \rangle}_{known} + \underbrace{\sqrt{(1 - \|z_q^{(\ell)}\|_2^2)(1 - \|z_c^{(\ell)}\|_2^2)}}_{known!\text{☺}} \\ &\quad \underbrace{\qquad\qquad}_{U_{q,c}^{\ell} \ \ upper\text{-}bound\ at\ level - \ell} \end{aligned}$$

6.2 ϵ-bounded Correctness Guarantee in Pyramid Rank

Pyramid Rank maintains thresholds $\tau_{\min}$ and $\tau_{\max}$ such that at termination, at least K items satisfy $U_{q,c_k}^L \geq \tau_{\min}$ and fewer than K satisfy $U_{q,c_k}^L \geq \tau_{\max}$ where $k \in \{1 \ldots K\}$. Because of the lower bound at termination in Pyramid Rank, $\tau_{\min} \leq \langle x_q^{(L)}, x_{c_k}^{(L)} \rangle$. Similarly, the upper bound at termination in Pyramid Rank ensures that all excluded items $j \notin \mathcal{I}$, $\langle x_q^{(L)}, x_{c_j}^{(L)} \rangle \leq \tau_{\max}$.

Now, the search terminates once $\tau_{\max} - \tau_{\min} \leq \epsilon$, which can be rewritten as $\tau_{\max} \leq \tau_{\min} + \epsilon$. Combining the upper bound, lower bound, and termination condition therefore we get, $\langle x_q^{(L)}, x_{c_j}^{(L)} \rangle \leq \tau_{\max} \leq \tau_{\min} + \epsilon \leq \langle x_q^{(L)}, x_{c_K}^{(L)} \rangle + \epsilon$, or in short:

$$\langle x_q^{(L)}, x_{c_j}^{(L)} \rangle \leq \langle x_q^{(L)}, x_{c_K}^{(L)} \rangle + \epsilon$$

This inequality guaranties that any unselected item's similarity score is within at most ϵ of the lowest-ranked retrieved item (c_K), providing a quantifiable retrieval quality bound of ϵ in Pyramid Rank. In other words, Pyramid Rank ensures that skipping full-resolution similarity computations for all database items using the longest possible embedding, leads to a bounded performance loss of at most ϵ which is a controllable hyperparameter.

MRL embedding	Foundation Model	NIGHTS ($i \to i$) R@1	R@5	R@10	nDCG@10
image (Kusupati et al., 2022)	✗	23.7	28.6	34.3	35.0
text (Zhang et al., 2025b)	✓	**29.5**	**36.0**	**58.4**	**39.8**

Table 10: Comparison of text- and image-based MRL embedding model for Pyramid Rank in image-to-image retrieval on the **NIGHTS** dataset. We report R@1, R@5, R@10, and nDCG@10. A checkmark (✓) indicates use of a foundation model for feature extraction, while a cross (✗) indicates its absence.

6.3 On the choice of MRL Embedding Model for Pyramid Rank

Foundation models provide semantically aligned, cross-modal embeddings that enable direct similarity computation between heterogeneous inputs to perform search and retrieval (Yuan, 2023). Foun-

dation embedding models are particularly efficacious for zero-shot retrieval across unseen modalities and domains without task-specific training (Li et al., 2024; Zhang et al., 2024c). For Pyramid Rank, we use the Qwen3-MRL 0.6B Embedding model (Zhang et al., 2025b), which is a foundation embedding model with large scale multimodal and multitask pretraining with support for pyramidal MRL (Kusupati et al., 2022) hierarchy. To the best of our knowledge, Qwen3-MRL 0.6B Embedding model (Zhang et al., 2025b) is the only embedding model that is both a foundation model and supports pyramidal MRL (Kusupati et al., 2022) hierarchy. In our experiments, we have a total of 5 cross-modal data types: i, (i, t), v, a, (a, v). However, Qwen3-MRL 0.6B Embedding model (Zhang et al., 2025b) operates on *text-only* input and therefore, we convert all data to text using captioning models. The original MRL study (Kusupati et al., 2022) offers a pre-trained ResNet (He et al., 2016) based image embedding model with support for pyramidal MRL embedding hierarchy. Despite being pretrained on ImageNet (Deng et al., 2009), ResNet (He et al., 2016) based image embedding model is not a foundation model. For $i \rightarrow i$ retrieval task, on NIGHTS (Fu et al., 2023) dataset, we compare both image based MRL embedding model and text based MRL embedding model for offline pre-processing in Pyramid Rank. In Tab. 10, we observe that the text-based MRL embedding (Zhang et al., 2025b), being derived from a foundation model, provides stronger retrieval performance compared to the image-based embedding (Kusupati et al., 2022), which lacks such foundation-level pretraining. It is important to note that our Pyramid Rank framework can be seamlessly integrated with any MRL-based embedding model, regardless of modality. In the next section, we discuss the offline data pre-processing and captioning strategy used to convert all multimodal data into a unified text representation prior to running Pyramid Rank.

6.4 Offline Data Pre-processing (Captioning) for Pyramid Rank

In Stage-1, all queries (processed at inference) and database items (pre-processed) are first converted to text via image/video captioning with Qwen3-VL-8B (Team, 2025b), audio transcription using Qwen2-Audio-7B (Chu et al., 2024); and audio + video data captioning using Qwen2.5-Omni-7B (Xu et al., 2025a). Across all 12 MMIR tasks, spanning 32 datasets, and both query and candidate modalities in our experiments, there are a total of 5 unique data modalities (combining both unimodal and cross-modal): i, (i, t), v, a, (a, v). This unified text conversion serves as a modality-agnostic pre-processing step, ensuring that all data, regardless of their original modality, can be encoded using a single foundation embedding model for consistent and comparable retrieval within Pyramid Rank.

Prompt for Caption Generation from Image with Qwen3-VL-8B (Team, 2025b) for queries in VisualNews, MSCOCO, Fashoin200K, Urban-1K, Flickr30K, NIGHTS and database candidates in VisualNews, MSCOCO, Fashoin200K, Urban-1K, Flickr30K, NIGHTS, FashoinIQ, CIRR, GenCIS:

Prompt for Caption Generation from Image:

```
You are given an image:      <i>
Generate a detailed caption describing
the key objects, their attributes, spatial
layout, and interactions.
Include information about the scene type,
context, and any salient visual cues that
convey intent or activity.
Output a detailed caption that captures the
essential meaning of the image.
```

Prompt for Caption Generation from Image + Text with Qwen3-VL-8B (Team, 2025b) for queries in OVEN, InfoSeek, FashoinIQ, CIRR, GenCIS and database candidates in EDIS, WebQA, OVEN, InfoSeek:

Prompt for Caption Generation from Image + Text:

```
You are given an image and its accompanying
text:      (<i>, <t>)
Generate a unified caption that integrates
both visual and textual information.
Describe how the image and text complement
each other.
Mention entities, actions, and context
shared across the image and text.
Output a detailed and coherent caption that
captures the essential, combined meaning of
the image and text.
```

Prompt for Caption Generation from Video with Qwen3-VL-8B (Team, 2025b) for queries in MSRVTT-1kA, MSVD, LSMDC, DiDeMo and database candidates in VMSRVTT-1kA, MSVD, LSMDC, DiDeMo:

Prompt for Caption Generation from Video:

```
You are given a short video clip:       <v>
Generate a temporally aware caption
describing the sequence of actions and
events.
Mention key objects, subjects, and
transitions over time, emphasizing movement
and interactions.
Output a detailed caption that captures the
essential meaning of the video.
```

Prompt for Caption Generation from Audio with Qwen2-Audio-7B (Chu et al., 2024) for database candidates in AudioCaps:

Prompt for Caption Generation from Audio:

```
You are given an audio clip:       <a>
Generate a descriptive caption summarizing
the content of the audio.
Include information about sound types,
speakers, emotions, acoustic events, and
temporal changes.
Output a detailed caption that captures the
essential meaning of the audio.
```

Prompt for Caption Generation from Audio + Video with Qwen2.5-Omni-7B (Xu et al., 2025a) for database candidates in AudioCaps:

Prompt for Caption Generation from Audio + Video:

```
You are given synchronized audio and video
data:       (<a>, <v>)
Generate a unified caption that combines
both auditory and visual cues.
Describe the main event or scene,
integrating spoken words, sounds, and
visible actions over time.
Output a detailed and coherent caption that
captures the essential, combined meaning of
the audio and video.
```

		image-text retrieval			
# CoT examples	CoT examples	R@1	R@5	R@10	nDCG@10
0	✗	51.3	57.9	63.5	66.0
2	✓	**57.5**	**65.8**	**74.2**	**67.1**

Table 11: Efficacy of using CoT examples in Bidirectional-CoT Embedding Score (S_{CoT}) used in Stage-2 re-ranking. Average Performance of image-text retrieval (averaged across 21 datasets) shown with and without Cot examples. We report R@1, R@5, R@10, and nDCG@10. A checkmark (✓) indicates use of CoT example, while a cross (✗) indicates its absence.

7 Stage-2: Additional Analysis

7.1 Bidirectional-CoT Embedding Score

In Tab. 11, Tab. 12, and Tab. 13, we report retrieval performance averaged across all 21 image-text retrieval datasets, 8 video-text retrieval datasets, and 3 audio-text retrieval datasets, respectively. The results in Tab. 11, Tab. 12, and Tab. 13 show that using CoT examples in Bidirectional-CoT Embedding Score significantly improves retrieval performance across modalities. Fig. 12 visually demonstrates how bidirectional Chain-of-Thought (CoT) prompting improves multimodal alignment between a query and the corresponding candidate in the Bidirectional-CoT Embedding Score. By explicitly guiding the model to reason about alignment before embedding extraction, the CoT-augmented setting (top) yields more contextually aware embeddings and higher cosine similarity compared to the non-CoT baseline (bottom), thereby enhancing fine-grained re-ranking in Stage-2.

		video-text retrieval			
# CoT examples	CoT examples	R@1	R@5	R@10	nDCG@10
0	✗	43.9	49.2	60.8	54.1
2	✓	**50.8**	**51.5**	**61.6**	**55.9**

Table 12: Efficacy of using CoT examples in Bidirectional-CoT Embedding Score (S_{CoT}) used in Stage-2 re-ranking. Average Performance of video-text retrieval (averaged across 8 datasets) shown with and without Cot examples. We report R@1, R@5, R@10, and nDCG@10. A checkmark (✓) indicates use of CoT example, while a cross (✗) indicates its absence.

		audio-text retrieval			
# CoT examples	CoT examples	R@1	R@5	R@10	nDCG@10
0	✗	31.5	36.3	39.7	40.4
2	✓	**39.4**	**41.9**	**44.8**	**42.1**

Table 13: Efficacy of using CoT examples in the proposed Bidirectional-CoT Embedding Score (S_{CoT}) used in Stage-2 re-ranking. Average Performance of audio-text retrieval (averaged across 3 datasets) shown with and without Cot examples. We report R@1, R@5, R@10, and nDCG@10. A checkmark (✓) indicates use of CoT example, while a cross (✗) indicates its absence.

7.2 Question Answering Relevance Score

In the example shown in Fig. 4, both candidate videos contain visually similar scenes involving water and human activity, which makes them difficult to distinguish. However, the Question Answering Relevance Score S_{QA} provides fine-grained discrimination by explicitly verifying the semantic attributes of the query through yes/no questions (e.g., *"Is the person swimming?"*, *"Are the rapids white water"*). While both candidates partially match the query, only C_2 correctly satisfies all query-specific attributes, leading to a higher S_{QA} score.

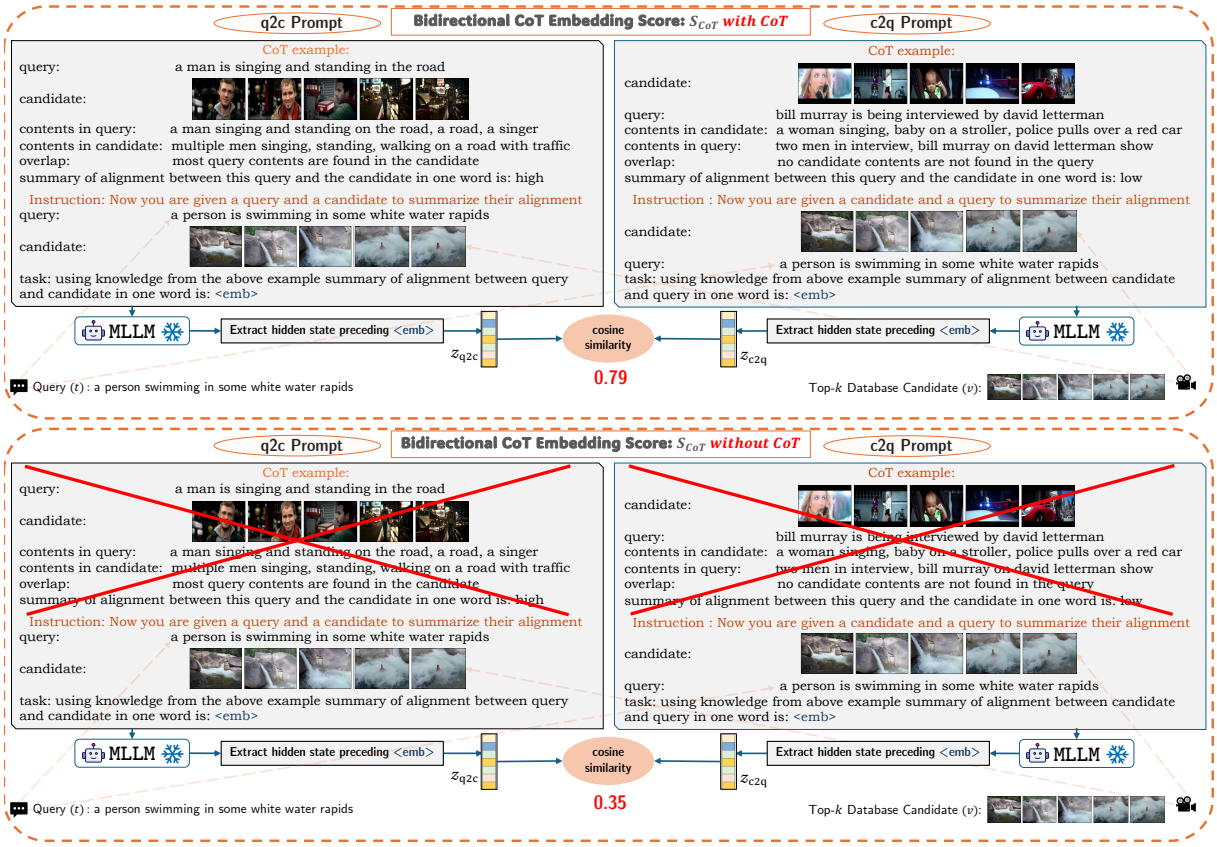

Figure 12: Visualization of the proposed Bidirectional-CoT Embedding Score (S_{CoT}) used in Stage-2 re-ranking. The upper panel illustrates the computation of S_{CoT} when incorporating Chain-of-Thought (CoT) examples, while the lower panel shows the same process without CoT. Incorporating CoT enables the MLLM to generate more semantically aligned embeddings, leading to higher similarity scores and improved retrieval quality.

Moreover, the intermediate QA logs serve as interpretable evidence for retrieval decisions, showing which visual or semantic aspects were matched or missed by each candidate; thus enhancing the explainability of Stage-2 re-ranking. Unlike C_2, C_1 includes water scenes but fails to meet key semantic conditions, such as swimming or visible rapids, lowering its score. The intermediate QA logs provide natural-language explanations that make this decision transparent; for instance, they explicitly record that C_1 answered "no" to "Is the person swimming?", whereas C_2 answered "yes" to all questions, thereby demonstrating how Stage-2 re-ranking uses interpretable QA to justify retrieval outcomes.

8 Additional Quantitative Analysis

8.1 Comparison with Additional Baselines on retrieval benchmarks

Tabs. 14 to 17 present an extended comparison of our MULTI-SCORE with additional baselines (both zero-shot and supervised) across a broad spectrum of multimodal retrieval tasks spanning text–image, image–text, text–video, and text–audio modalities.

Our MULTI-SCORE consistently achieves state-of-the-art zero-shot results across the majority of datasets within the M-BEIR (Wei et al., 2024) benchmark (Tab. 14), as well as on additional text–image retrieval benchmarks (Tab. 15). Specifically, the proposed Bidirectional-CoT Embedding Score and Question Answering Relevance Score re-ranking modules help achieve strong retrieval accuracy even in challenging cross-modal tasks, outperforming prior multimodal large language model (MLLM) based retrieval systems as well as dual encoder based retrieval systems.

8.2 Zero-shot Multi-Score often outperforms supervised audio-video retrieval benchmarks

Tabs. 16 and 17 shows that MULTI-SCORE, despite being zero-shot often outperforms supervised audio-video retrieval benchmarks. Audio and video retrieval require rich cross-modal understanding such as temporal, compositional aspects of such

	Methods	$t \to i$			$t \to t$	$t \to (i,t)$		$i \to t$			$i \to i$	$(i,t) \to t$		$(i,t) \to i$		$(i,t) \to (i,t)$	
		VisualNews R@5	MSCOCO R@5	Fashion200K R@10	WebQA R@5	EDIS R@5	WebQA R@5	VisualNews R@5	MSCOCO R@5	Fashion200K R@10	NIGHTS R@5	OVEN R@5	InfoSeek R@5	FashionIQ R@10	CIRR R@5	OVEN R@5	InfoSeek R@5
Zero-Shot	CLIP-L (Radford et al., 2021)	43.3	61.1	6.6	36.2	43.3	45.1	41.3	79.0	7.7	26.1	24.2	20.5	7.0	13.2	38.8	26.4
	SigLIP (Zhai et al., 2023)	30.1	75.7	36.5	39.8	27.0	43.5	30.8	88.2	34.2	28.9	29.7	25.1	14.4	22.7	41.7	27.4
	BLIP (Li et al., 2022)	16.4	74.4	15.9	44.9	26.8	20.3	17.2	83.2	19.9	27.4	16.1	10.2	2.3	10.6	27.4	16.6
	BLIP2 (Li et al., 2023)	16.7	63.8	14.0	38.6	26.9	24.5	15.0	80.0	14.2	25.4	12.2	5.5	4.4	11.8	27.3	15.8
	Qwen2-VL-7B (Wang et al., 2024)	9.3	55.1	5.0	42.0	26.2	9.4	5.4	46.6	4.0	21.3	21.4	22.5	4.3	16.3	43.6	36.2
	Qwen2.5-VL-3B (Team, 2025a)	10.8	56.7	6.8	43.9	27.4	10.9	6.2	48.3	5.6	22.1	23.2	23.6	3.1	15.1	42.2	34.8
	Qwen2.5-VL-7B (Team, 2025a)	11.9	55.4	7.1	45.2	26.3	12.4	7.7	47.1	6.4	23.7	24.0	22.2	2.6	16.0	43.5	33.9
	Qwen3-VL-4B (Team, 2025b)	12.4	56.1	8.2	44.1	27.5	11.8	8.6	45.9	7.2	24.9	22.7	23.4	3.5	14.9	44.7	35.1
	Qwen3-VL-8B (Team, 2025b)	15.8	59.0	11.7	46.0	30.9	13.6	10.1	49.5	9.0	27.8	26.0	24.8	6.8	16.2	48.4	37.0
	Qwen3-Omni-30B-Th (Xu et al., 2025b)	16.6	59.9	12.5	46.9	31.7	14.6	10.9	50.5	9.7	28.6	26.7	25.7	7.5	17.0	49.2	37.9
	Qwen3-MRL (Kusupati et al., 2022)	32.5	57.2	19.0	50.3	39.8	47.1	20.4	55.0	15.7	32.7	28.5	26.3	25.4	21.2	54.0	42.8
	MULTI-SCORE	**49.8**	**86.9**	**41.8**	**77.4**	**79.3**	**81.1**	**49.7**	**92.3**	**38.4**	**36.0**	**50.4**	**62.7**	**38.5**	**58.3**	**68.7**	**54.9**
Supervised	UniR-BLIP$_{FF}$ (Wei et al., 2024)	23.4	79.7	26.1	80.0	50.9	79.8	22.8	89.9	28.9	33.0	41.0	22.4	29.2	52.2	55.8	33.0
	UniR-CLIP$_{SF}$ (Wei et al., 2024)	42.6	81.1	18.0	84.7	59.4	78.7	43.1	**92.3**	18.3	32.0	45.5	27.9	24.4	44.6	67.6	48.9
	GENIUS (Kim et al., 2025)	18.5	68.1	13.7	32.5	37.0	49.7	18.7	83.2	12.8	8.2	36.6	11.2	13.2	20.7	19.3	39.5
	GENIUS^{R} (Kim et al., 2025)	27.4	78.0	16.2	44.6	44.3	60.6	28.4	91.1	16.3	30.2	41.9	20.7	19.3	39.5	52.5	30.1
	LamRA-ret (Liu et al., 2025)	41.6	81.5	28.7	86.0	62.6	81.2	39.6	90.6	30.4	32.1	54.1	52.1	33.2	53.1	76.2	63.3
	LamRA (Liu et al., 2025)	**48.0**	**85.2**	**32.9**	**96.7**	**75.8**	**87.7**	**48.6**	**92.3**	**36.1**	**33.5**	**59.2**	**64.1**	**37.8**	**63.3**	**79.2**	**78.3**

Table 14: Results on the M-BEIR benchmark (Wei et al., 2024) with each dataset and its standard evaluation metric. MULTI-SCORE achieves the best zero-shot performance on all M-BEIR tasks while occasionally outperforming some stat-of-the-art supervised baselines. Colors: best zero-shot results: **bold blue**; best supervised results: **bold red**; best overall results: highlighted yellow.

	Methods	$t \to i$		$i \to t$		$(i,t) \to i$
		Urban-1K	Flickr30K	Urban-1K	Flickr30K	GeneCIS
Zero-Shot	CLIP-L (Radford et al., 2021)	52.8	67.3	68.7	87.2	13.3
	Long-CLIP-L (Zhang et al., 2024a)	86.1	76.1	82.7	89.3	16.3
	UniIR-CLIP (Wei et al., 2024)	75.0	78.7	78.4	94.2	16.8
	E5-V (Jiang et al., 2024b)	84.0	79.5	82.4	88.2	18.5
	MagicLens-L (Zhang et al., 2024b)	59.3	72.5	24.2	84.6	16.3
	EVA-CLIP-18B (Sun et al., 2024)	81.7	83.3	83.3	96.7	13.6
	Qwen2-VL-7B (Wang et al., 2024)	53.0	61.9	53.3	64.8	11.7
	Qwen2.5-VL-3B (Team, 2025a)	53.8	61.3	52.8	65.7	12.4
	Qwen2.5-VL-7B (Team, 2025a)	54.6	62.2	52.2	66.6	13.0
	Qwen3-VL-4B (Team, 2025b)	55.1	61.6	53.0	67.7	14.1
	Qwen3-VL-8B (Team, 2025b)	56.0	62.4	53.9	67.1	14.9
	Qwen3-Omni-30B-Th (Xu et al., 2025b)	56.9	62.0	54.6	68.0	14.6
	Qwen3-MRL (Kusupati et al., 2022)	67.3	69.4	63.6	73.9	19.5
	MULTI-SCORE	**98.2**	**89.6**	**97.8**	**98.1**	**34.6**
Sup.	LamRA-ret (Liu et al., 2025)	95.1	82.8	94.3	92.7	18.9
	LamRA (Liu et al., 2025)	**98.8**	**88.1**	**98.0**	**97.6**	**24.8**

Table 15: Additional results on image-text retrieval tasks (R@1). Colors: best zero-shot results: **bold blue**; best supervised results: **bold red**; best overall results: highlighted yellow.

	Methods	$t \to a$		$t \to (a,v)$
		AudioCaps R@1	Clotho R@1	AudioCaps R@1
Zero-Shot	AVR-VC3M (Nagrani et al., 2022)	8.7	10.6	3.0
	Qwen2-Audio-7B (Chu et al., 2024)	29.7	18.3	23.5
	Qwen2.5-Omni-7B (Xu et al., 2025a)	31.1	18.5	30.9
	Qwen3-Omni-30B-Th (Xu et al., 2025b)	32.1	15.4	21.8
	Qwen3-MRL (Kusupati et al., 2022)	35.4	17.9	22.7
	MULTI-SCORE	**45.4**	**28.2**	**44.5**
Sup.	SOTA (Oncescu et al., 2021)	24.3	6.7	28.1
	AVR (Nagrani et al., 2022)	32.0	7.8	41.4
	AVR-VC3M (Nagrani et al., 2022)	**35.5**	**8.4**	**43.2**

Table 16: Performance comparison for cross-modal text-audio retrieval tasks in terms of R@1. Colors: best zero-shot results: **bold blue**; best supervised results: **bold red**; best overall results: highlighted yellow.

data that is often weakly supervised or poorly captured by fixed joint embeddings. Our two-stage framework first preserves semantic similarity through coarse MRL-based efficient filtering, then leverages MLLM-based QA re-ranking to perform fine-grained temporal and contextual alignment, allowing MULTI-SCORE to better align complex audio–visual temporal content with queries than task-specific supervised models.

Overall, these results reaffirm the robustness, scalability, and modality-agnostic generalization of MULTI-SCORE as a unified zero-shot multimodal retrieval framework that efficiently achieves state-of-the-art performance across visual, auditory, textual, and cross-modal tasks.

8.3 Generalization to Unseen Retrieval Tasks

Tab. 18 reports retrieval performance (R@5) on the M-BEIR (Wei et al., 2024) benchmark, evaluating generalization to *unseen retrieval tasks*. Our proposed zero-shot MULTI-SCORE is compared against state-of-the-art **Zero-Shot** and **Supervised** baselines. The supervised baselines are trained on five of the eight M-BEIR tasks, while evaluation is performed on the three held-out tasks to evaluate task generalization capabilities.

Despite the supervised methods being fine-tuned on related tasks, MULTI-SCORE achieves superior or comparable performance across all unseen settings $i \to i, (i,t) \to t, (i,t) \to (i,t)$, with an average R@5 of **54.5**, outperforming both zero-shot and supervised baselines. This demonstrates that our two-stage framework, particularly the combination of Bidirectional-CoT Embedding Score and Question Answering Relevance Score enables strong cross-domain generalization without any task-specific fine-tuning. Overall, MULTI-SCORE exhibits robust *generalization to unseen retrieval tasks*, highlighting its scalability and adaptability as a universal zero-shot multimodal retrieval system.

8.4 Performance on Additional Evaluation Metrics

Tab. 23 reports the retrieval performance of MULTI-SCORE across all 12 MMIR tasks in terms

	Methods	$t \rightarrow v$				$v \rightarrow t$			
		MSRVTT-1kA R@1	MSVD R@1	LSMDC R@1	DiDeMo R@1	MSRVTT-1kA R@1	MSVD R@1	LSMDC R@1	DiDeMo R@1
Zero-Shot	NoiseEst (Amrani et al., 2021)	8.0	13.7	4.2	–	–	–	–	–
	Frozen (Bain et al., 2021)	18.7	33.7	9.3	21.1	–	–	–	–
	MCQ-BridgeFormer (Ge et al., 2022)	26.0	43.6	12.2	25.6	–	–	–	–
	OA-Trans (Wang et al., 2022)	23.4	–	–	23.5	–	–	–	–
	LamRA (Liu et al., 2025)	44.7	52.4	–	–	–	–	–	–
	Qwen2.5-VL-3B (Team, 2025a)	39.5	51.2	21.2	37.1	48.0	27.5	59.6	34.0
	Qwen2.5-VL-7B (Team, 2025a)	37.2	48.8	19.6	35.4	46.9	25.7	59.2	32.2
	Qwen3-VL-4B (Team, 2025b)	38.2	49.5	20.0	36.3	46.6	26.4	60.3	32.8
	Qwen3-VL-8B (Team, 2025b)	38.6	50.4	20.6	36.1	47.8	27.0	61.2	33.3
	Qwen3-Omni-30B-Th (Xu et al., 2025b)	39.5	51.0	21.1	37.3	48.5	26.8	62.2	34.1
	MULTI-SCORE	**55.7**	**69.4**	**27.1**	**50.3**	**53.1**	**31.9**	**68.0**	**50.7**
Supervised	NoiseEst (Amrani et al., 2021)	17.4	20.3	6.4	–	–	–	–	–
	Frozen (Bain et al., 2021)	31.0	45.6	15.0	31.0	–	–	–	–
	MCQ-BridgeFormer (Ge et al., 2022)	37.6	**52.0**	17.9	37.0	–	–	–	–
	OA-Trans (Wang et al., 2022)	35.8	39.1	18.2	34.8	17.5	–	–	–
	T2VLAD (Wang et al., 2021)	29.5	–	14.3	–	31.8	–	14.2	–
	X-Pool (Gorti et al., 2022)	46.9	47.2	**25.2**	–	43.9	**66.4**	**22.7**	–
	RPC (Lai et al., 2025)	47.3	38.5	22.8	34.7	46.3	48.1	22.0	33.4
	NeighborRetr (Lin et al., 2025)	**49.5**	47.9	–	**48.2**	**48.7**	63.3	–	**48.4**

Table 17: Performance comparison for cross-modal text-video retrieval tasks in terms of R@1. Colors: best zero-shot results: **bold blue**; best supervised results: **bold red**; best overall results: highlighted yellow.

	Methods	$i \rightarrow i$	$(i,t) \rightarrow t$		$(i,t) \rightarrow (i,t)$		Average
		NIGHTS	OVEN	InfoS	OVEN	InfoS	
Zero-Shot	LamRA-Ret (Liu et al., 2025)	27.2	44.7	44.0	62.8	49.5	45.6
	LamRA (Liu et al., 2025)	29.2	46.9	54.2	65.1	59.1	50.9
	MULTI-SCORE	**36.0**	**50.4**	**62.7**	**68.7**	**54.9**	**54.5**
Supervised	UniR-BLIP$_{FF}$ (Wei et al., 2024)	**33.0**	41.0	22.4	55.8	33.0	37.0
	UniR-CLIP$_{SF}$ (Wei et al., 2024)	32.0	**45.5**	**27.9**	**67.6**	**48.9**	**44.4**

Table 18: Retrieval performance (R@5) on the M-BEIR dataset (Wei et al., 2024), evaluating generalization to unseen retrieval tasks. Our proposed Zero-Shot MULTI-SCORE is compared against state-of-the-art Zero-Shot and Supervised baselines. The Supervised models are trained on the remaining five tasks within M-BEIR. Even though the Supervised baselines are trained on other retrieval tasks, MULTI-SCORE demonstrates superior unseen task generalization, achieving higher retrieval performance on unseen tasks without any fine-tuning. Colors: best zero-shot results: **bold blue**; best supervised results: **bold red**; best overall results: highlighted yellow.

of R@1, R@5, R@10, and nDCG@10. While Recall@K (**R@K**) measures the proportion of relevant items retrieved within the top-K results, emphasizing the ability of the model to correctly identify relevant candidates, normalized Discounted Cumulative Gain (**nDCG@K**) on the other hand, accounts for the ranking order of those retrieved items, assigning higher importance to items that appear both correctly and earlier in the ranking. Thus, R@K captures retrieval *coverage*, and nDCG@K reflects retrieval *quality*.

Overall, MULTI-SCORE achieves strong performance across all metrics and modalities, demonstrating that the proposed two-stage framework not only retrieves relevant candidates efficiently (high Recall) but also maintains their correct relative ordering (high nDCG). The consistently high nDCG values across all 12 MMIR tasks indicate that the reranking stage effectively improves semantic alignment, confirming the robustness and fine-grained discriminative capability of Bidirectional-CoT Embedding Score and Question Answering Relevance Score.

9 Detailed Compute Cost Analysis for Multi-Score

Modality	Model	Avg time per item
Image Captioning	Qwen3-VL-8B	0.18 s
Video Captioning	Qwen2.5-Omni-7B	0.65 s
Audio Captioning	Qwen2-Audio-7B	0.35 s

Table 19: Offline pre-processing (Captioning) cost analysis in MULTI-SCORE.

Offline Pre-processing (One-Time Cost). Indexing the full 5.7M-item database requires approximately 21 hours on a 32-GPU cluster and is performed once offline; no captioning or embedding extraction occurs at online inference or query time.

Modality	Model	Avg time per item
Text Embedding	Qwen3-MRL	0.012 s

Table 20: Offline pre-processing (Embedding Extraction) cost analysis in MULTI-SCORE.

Database Size (N)	naïve MRL Inference time (ms)	naïve MRL R@100	Pyramid Rank Inference time (ms)	Pyramid Rank R@100	Inference Speedup	Recall Preservation Ratio
100K	3.2	86.4	1.8	86.3	1.8x	1.00
500K	15.7	86.4	6.3	86.3	2.5x	1.00
1M	31.4	86.4	11.5	86.3	2.7x	1.00
5.7M	179.2	86.4	54.6	86.2	3.3x	1.00

Table 21: Stage-1: Online Pyramid Rank cost for universal retrieval (32 datasets combined).

The offline captioning cost breakdown is shown in Tab. 19. The offline embedding extraction cost breakdown is shown in Tab. 20. Total database size: 5.7M items, Embedding time per item: 0.012 s. Total offline embedding extraction cost (parallelized across 32 GPUs) = 5.7M $\times$ 0.12 / 32 = 0.6 GPU-hours. Total offline pre-processing cost = captioning+embedding extraction = 20 + 0.6 = 20.6 GPU-hours.

Stage-2 Component	MLLM Re-ranking (Per Query) Inference Time (ms)
Bidirectional-CoT Embedding Score	6.5
Question Answering Relevance Score	8.2

Table 22: Stage-1: Our proposed Stage-2 Online MLLM Re-ranking cost for universal retrieval (32 datasets combined).

Online Inference (Per Query Cost). Tab. 21 shows Pyramid Rank (Stage-1) achieves up to 3.3× inference speedup at 5.7M scale while nearly maintaining identical R@100, demonstrating that efficiency gains do not compromise retrieval quality. In Stage-2, Bidirectional-CoT Embedding Score and Question Answering Relevance Score are run in parallel per query using distributed inference and KV caching. Bidirectional-CoT Embedding Score cost include time for 2 forward passes (parallel), and time to extract hidden state before <emb> token. Question Answering Relevance Score cost include QA with 5 questions, maximun of 5 tokens per answer. Tab. 22 shows Bidirectional-CoT Embedding Score and Question Answering Relevance Score running costs. Total Stage-2 time (K=100) K $\times$ max(6.5, 8.2) = 100 $\times$ 8.2 = 0.82 s. Total Per-Query Online Inference (5.7M Database) Stage (1+2) = 54.6 ms + 0.82 s = 0.87 s.

Task	Dataset	R@1	R@5	R@10	nDCG@10
$t \to i$	VisualNews	46.5	49.8	67.4	53.7
	MSCOCO	82.7	86.9	91.5	86.2
	Fashion200K	32.7	37.7	41.8	36.5
	Urban-1K	98.2	98.5	98.9	99.0
	Flickr30K	89.6	92.1	93.6	91.5
$t \to t$	WebQA	50.9	77.4	85.3	66.4
$t \to (i,t)$	EDIS	52.5	79.3	86.5	67.9
	WebQA	60.4	81.1	87.7	72.6
$i \to t$	VisualNews	45.4	49.7	67.3	53.1
	MSCOCO	87.3	92.3	95.0	90.6
	Fashion200K	34.2	40.2	38.4	38.9
	Urban-1K	97.8	98.0	98.0	95.7
	Flickr30K	98.1	98.4	98.7	97.9
$i \to i$	NIGHTS	29.5	36.0	58.4	39.8
$(i,t) \to t$	OVEN	37.9	50.4	67.8	49.6
	InfoSeek	46.3	62.7	75.8	58.5
$(i,t) \to i$	FashionIQ	30.3	34.8	38.5	33.7
	CIRR	45.9	58.3	72.9	56.6
	GeneCIS	34.6	40.8	46.1	39.3
$(i,t) \to (i,t)$	OVEN	40.6	68.7	79.7	57.9
	InfoSeek	34.7	54.9	70.0	49.6
$t \to v$	MSRVTT-1kA	55.7	61.2	84.4	65.8
	MSVD	69.4	73.5	81.9	74.1
	LSMDC	27.1	31.4	38.0	31.3
	DiDeMo	50.3	54.8	56.1	52.9
$v \to t$	MSRVTT-1kA	53.1	61.7	81.6	63.7
	MSVD	31.9	40.1	43.7	37.1
	LSMDC	68.0	72.5	80.8	72.9
	DiDeMo	50.7	55.9	62.7	55.4
$t \to a$	AudioCaps	45.4	47.8	53.4	48.4
	Clotho	28.2	32.9	33.0	30.5
$t \to (a,v)$	AudioCaps	44.5	45.0	48.1	45.7

Table 23: Retrieval Performance of MULTI-SCORE for all 12 MMIR tasks in terms of R@1 (↑), R@5 (↑), R@10 (↑), nDCG@10 (↑).

10 Additional Qualitative Analysis

Fig. 13 compares qualitative retrieval outcomes across MULTI-SCORE (ours), MULTI-SCORE (naïve MRL), and the supervised LamRA baseline for a query involving fine-grained visual understanding. Our two-stage framework efficiently filters and re-ranks candidates, retrieving the correct target within 0.09 s, substantially faster and more semantically accurate than competing methods. While the naïve MRL variant retrieves visually similar but less precise candidates and incurs higher latency (0.29 s), LamRA achieves moderate accuracy but still relies on supervised alignment. These results highlight MULTI-SCORE's capability to achieve fine-grained and efficient retrieval without supervision.

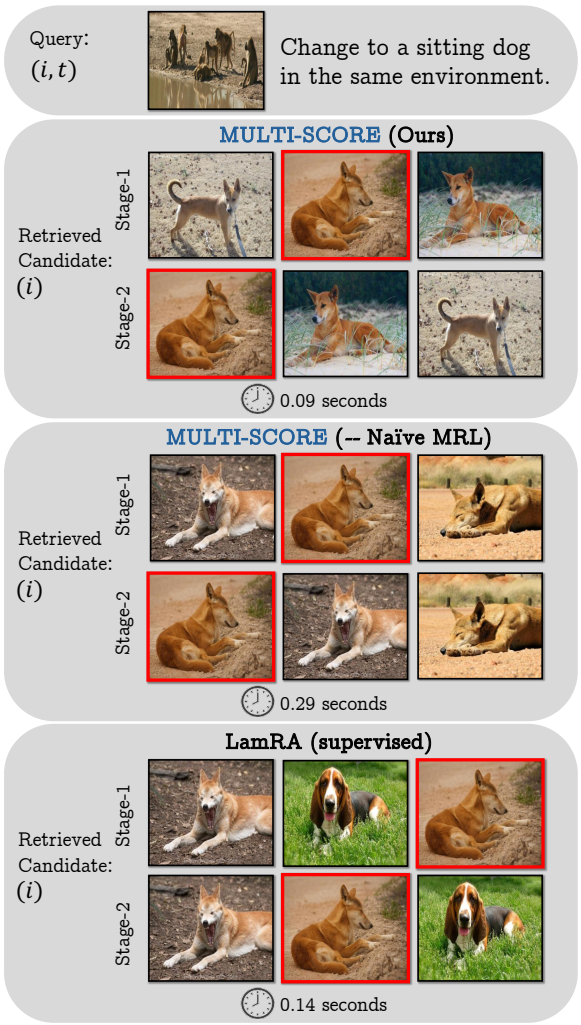

Figure 13: Qualitative Results: Given a multimodal query that combines an image and a text instruction, MULTI-SCORE retrieves the correct and semantically aligned target (red box) faster and more accurately than the Naïve MRL and LamRA baselines. Unlike LamRA, which misinterprets scene context (e.g., retrieving images from mismatched environments), MULTI-SCORE effectively grounds textual edits within the visual context, highlighting its ability to perform precise and context-aware multimodal retrieval. While MULTI-SCORE correctly grounds the query in the visual environment to rank relevant results, LamRA misinterprets the scene context, retrieving images with incorrect (i.e., grassy) environments.