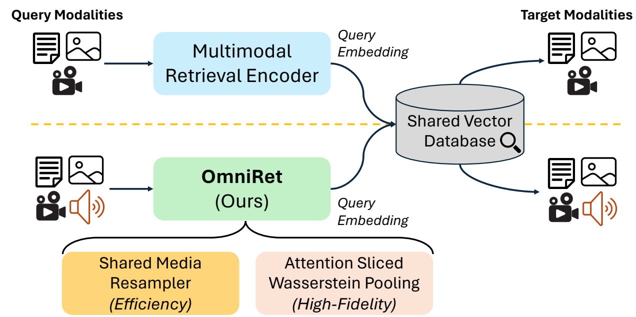
Efficient and High-Fidelity Omni Modality Retrieval

Chuong Huynh[1†] Manh Luong[2†] Abhinav Shrivastava[1]
[1] University of Maryland, College Park, USA [2] Monash University, Australia [†] Equal contribution

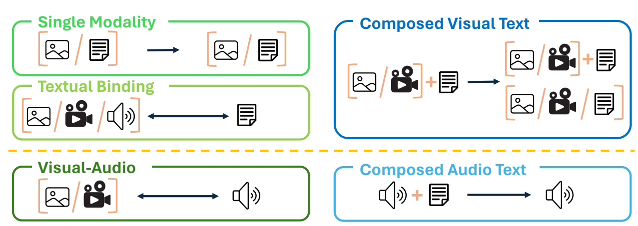
Project Page: https://hmchuong.github.io/omniret

Abstract

Multimodal retrieval is the task of aggregating information from queries across heterogeneous modalities to retrieve desired targets. State-of-the-art multimodal retrieval models can understand complex queries, yet they are typically limited to two modalities: text and vision. This limitation impedes the development of universal retrieval systems capable of comprehending queries that combine more than two modalities. To advance toward this goal, we present OmniRet, the first retrieval model capable of handling complex, composed queries spanning three key modalities: text, vision, and audio. Our OmniRet model addresses two critical challenges for universal retrieval: computational efficiency and representation fidelity. First, feeding massive token sequences from modality-specific encoders to Large Language Models (LLMs) is computationally inefficient. We therefore introduce an attention-based resampling mechanism to generate compact, fixed-size representations from these sequences. This shared module is designed to maintain representational diversity and generalization capabilities while remaining sensitive to modality-specific information. Second, compressing rich omni-modal data into a single embedding vector inevitably causes information loss and discards fine-grained details. We propose Attention Sliced Wasserstein Pooling to preserve these fine-grained details, leading to improved omni-modal representations. OmniRet is trained on an aggregation of approximately 6 million query-target pairs spanning 30 datasets. We benchmark our model on 13 retrieval tasks and a MMEBv2 subset. Our model demonstrates significant improvements on composed query, audio and video retrieval tasks, while achieving on-par performance with state-of-the-art models on others. Furthermore, we curate a new Audio-Centric Multimodal Benchmark (ACM). This new benchmark introduces two critical, previously missing tasks—composed audio retrieval and audio-visual retrieval—to more comprehensively evaluate a model's omni-modal embedding capacity. We believe our benchmark will facilitate the development of universal retrieval systems.



(a) Multimodal Retrieval Systems (top) and our OmniRet (bottom)



(b) Our evaluation suite with new Audio-Centric MultiModal tasks.

Figure 1. **(a) Comparison of retrieval systems.** OmniRet is the first to handle composed queries from text, vision, and audio. It introduces a Shared Media Resampler for efficiency and Attention Sliced Wasserstein Pooling to preserve high-fidelity, fine-grained details. **(b) Our Audio-Centric MultiModal evaluation suite.** We introduce two novel, audio-centric tasks: Visual-Audio Retrieval and Composed Audio-Text Retrieval. These additions fill critical gaps in existing benchmarks, enabling a more comprehensive evaluation of omni-modal retrieval models.

1. Introduction

The landscape of information retrieval has fundamentally shifted from single-modality systems, such as traditional text search, to a complex ecosystem demanding seamless interaction across diverse data types. The proliferation of digital content requires models that can understand heterogeneous, composed queries—for example, finding a target image using a query that combines a related image and a line of text. While pioneering cross-modal models like CLIP [57], BLIP [36, 37], and CLAP [74] have excelled at bi-modal tasks, they fall short in scenarios requiring compositional understanding across multiple modalities. Furthermore, the frontier of omni-modal retrieval, often relying on Multimodal LLMs (MLLMs) [27, 41, 46], reveals significant gaps. Modalities like video and, most notably, audio remain critically underserved due to a scarcity of dedicated models and comprehensive training data.

The path to a universal retrieval model is obstructed by two primary challenges: *representation fidelity* and *computational efficiency*. On one hand, condensing rich, multimodal inputs into a single embedding vector creates an information bottleneck that discards the fine-grained details necessary for high-fidelity retrieval [72]. Prior works [27, 41] often resort to simple average pooling or using the '[EOS]' token as the final representation. These methods are efficient, however, the fine-grained information extracted from LLMs could be discarded due to the weighted sum operation. While sophisticated late-interaction mechanisms like ColBERT [28] overcome this by retaining token-level embeddings, they introduce prohibitive computational and storage costs, making them impractical for large-scale systems. On the other hand, the sheer scale of modern foundation models presents a significant efficiency challenge. Feeding massive token sequences from media encoders, often exceeding 500 tokens for a single image, into an LLM composer results in an explosion of computational cost. This computational overhead severely limits the feasible batch size, which in turn undermines the effectiveness of contrastive learning, a cornerstone for training state-of-the-art embedding models.

To address these challenges, we introduce OmniRet, the first unified framework for omni-modal embedding, capable of handling multimodal queries and targets from three primary modalities: text, vision, and audio (Fig. 1a). Our framework makes two primary contributions. First, to tackle the efficiency bottleneck, we introduce an attention-based resampling module that intelligently condenses large sequences of media tokens into a compact, fixed-size set. This module is shared across modalities to enhance generalization, yet maintains modality-specific sensitivity by using separate media latents. We augment this with a novel diversity loss function, ensuring that these condensed tokens retain maximal information and thereby improve the final embedding quality. Second, to achieve high-fidelity representation without sacrificing efficiency, we propose a new pooling method called Attention Sliced Wasserstein Pooling (ASWP), inspired by PSWE [52]. By conceptualizing the set of LLM output tokens as a distribution, ASWP computes a rich embedding based on the distance to a set of learnable references [53]. This approach preserves fine-grained, token-level information while maintaining the speed and simplicity of a single-vector system. Our extensive experiments demonstrate that OmniRet achieves superior performance on composed retrieval tasks, including those for images, videos, and audio.

We also introduce a new benchmark, the Audio-Centric Multimodal (ACM) benchmark (Fig. 1b). This benchmark is curated from the VGG-Sound dataset to evaluate universal retrieval systems that extend beyond the traditional text-vision paradigm. We leverage powerful generative models (QwenOmni2.5 [76] and Gemini2.5 [9]) to generate high-quality audio captions and modification text describing the relationship between audio pairs. We validate the quality of our synthetic benchmark via human evaluation. Our benchmark contributes two major, previously missing retrieval tasks: composed audio retrieval and audio-visual retrieval.

Our main contribution can be summarized as: (i) We introduce OmniRet, the first universal retrieval framework capable of handling text, vision, and audio. OmniRet addresses the critical challenges of computational efficiency and representation fidelity through a novel attention-based resampler and our Attention Sliced Wasserstein Pooling (ASWP) technique. (ii) We propose a large-scale, multi–task training strategy to train OmniRet on a massive aggregation of approximately 6 million query-target pairs from 30 distinct datasets. (iii) We curate and will release the ACM benchmark, a new audio-centric benchmark featuring two novel tasks, to advance the evaluation of universal retrieval systems. (iv) Through extensive experiments, we demonstrate that OmniRet achieves state-of-the-art performance, particularly on composed retrieval tasks across image, video, and audio.

2. Related Works

Multimodal Embedding. Recent advances in multimodal embedding have been driven by large-scale paired datasets and self-supervised learning. Pioneering cross-modal models such as CLIP variants [24, 57, 63] and BLIP [36, 37] focus on text-vision alignment, while others like CLAP [12, 48] target text-audio modalities. Beyond dataset-centric improvements, several works explore improved pretraining objectives for learning better joint embedding spaces, including SigLIP variants [66, 78] and AlignCLIP [13]. More recently, the emergence of LLMs has motivated approaches that leverage their strong representational power [17, 69] to construct unified embedding spaces.

Universal Multimodal Retrieval. Research on universal retrieval systems has rapidly advanced. UniIR [71] pioneered a universal retriever by training a dual-encoder on 20 datasets, establishing a strong any-to-any benchmark; however, it relies on separate, specialized encoders. Other approaches adapt existing VLMs for embedding tasks [25, 26, 50, 81], achieving strong performance but often limits to text-vision modalities. To improve retrieval, some models [41, 46] leverage the instruction-following ability of LLMs, using generative rewards [80, 82] to refine predictions. This line of work is complemented by advances in data curation [22, 79, 82] to source high-quality cross-modal pairs for training. Moving beyond two modalities, ImageBind [16] demonstrated the feasibility of a joint embedding space for six modalities. While this enables omni-retrieval, it highlights a key challenge: managing massive, high-dimensional media inputs from varied encoders. These methods often remain computationally inefficient by directly processing all encoder tokens. Our work directly addresses this efficiency bottleneck.

Embedding Pooling. Embedding pooling techniques are critical for aggregating a sequence of embedding vectors into a single, fixed-size vector [55]. Simple methods, like mean pooling or using the '[EOS]' token's hidden state, are fast but often suboptimal, discarding fine-grained information. More advanced techniques, such as NV-Embed [33], use learnable queries to generate a more descriptive single-vector representation. However, the entire single-vector paradigm faces theoretical limitations [72], spurring the development of late-interaction models. ColBERT [28] exemplifies this approach by performing expensive but highly effective token-level comparisons. Subsequent research has focused on making this paradigm more efficient, such as MetaEmbed [75]. While these advancements make late interaction more feasible, they still deviate from the single-vector format required for highly optimized Approximate Nearest Neighbor indexes. To bridge this gap, we introduce a new multimodal-aggregation technique, ASWP, inspired by PSWE [52]. Our method captures fine-grained relevance while remaining fully compatible with efficient, large-scale retrieval systems.

3. OmniRet

3.1. Problem Definition

Omni-modal information retrieval aims to match a query $\mathbf{q}$ to a set of candidates $\mathbf{c}$, where both can be of any modality. To handle diverse retrieval tasks, we prepend a natural language instruction to the query:

$$\mathbf{q}_{\text{inst}} = \text{Instruct: } \{\textit{task\_definition}\} \; \backslash n \text{ Query: } \{\mathbf{q}\} \tag{1}$$

where '$\{$*task_definition*$\}$' specifies the retrieval goal and domain. This instruction-tuning approach enhances the model's generalization capabilities. The candidates $\mathbf{c}$, however, remain in their raw, instruction-free form.

Our goal is to train a unified embedding model, f, that maps both the instructed query and the candidate to a shared D-dimensional space: $\mathbf{h^q} = f(\mathbf{q}_{\text{inst}})$ and $\mathbf{h^c} = f(\mathbf{c})$. We train f using a Hard-negative InfoNCE loss [56] over in-batch negatives:

$$\mathcal{L}_{\text{cont}} = -\log \frac{e^{\phi(\mathbf{h^q}, \mathbf{h^{c+}})}}{\sum_c w(\mathbf{h^q}, \mathbf{h^c}) \cdot e^{\phi(\mathbf{h^q}, \mathbf{h^c})}} \tag{2}$$

where $\phi(\mathbf{x}, \mathbf{y}) = \frac{1}{\tau}\cos(\mathbf{x}, \mathbf{y})$ is the temperature-scaled (τ) cosine similarity score. Here, $\mathbf{c}+$ and $\mathbf{c}-$ denote positive and negative candidates, respectively. $w(\cdot)$ is an adaptive weighting term for hard-negative mining where $w(\mathbf{h^q}, \mathbf{h^{c+}}) = 1$ and:

$$w(\mathbf{h^q}, \mathbf{h^{c-}}) = \frac{|\mathcal{N}| e^{\beta\phi(\mathbf{h^q}, \mathbf{h^{c-}})}}{\sum_{c-} e^{\beta\phi(\mathbf{h^q}, \mathbf{h^{c-}})}} \tag{3}$$

with $|\mathcal{N}|$ is the number of negative samples. Following prior works, we set the temperature $\tau = 0.07$ and weighting parameter $\beta = 0.5$.

To further providing a strong discriminative learning signal to the model, we also adopt the hinge-based triplet loss

$$\mathcal{L}_{\text{triplet}} = \sum_{c-} \max[\eta + \phi(\mathbf{h^q}, \mathbf{h^{c-}}) - \phi(\mathbf{h^q}, \mathbf{h^{c+}})] \tag{4}$$

where $\eta = 0.1$ is the margin hyper-parameter.

3.2. Model Architecture

Our model, OmniRet, employs a Large Language Model (LLM) as a universal composer to process inputs from various modalities, as depicted in Fig. 2. Text inputs are processed directly by the LLM, while other modalities are first encoded by specialized encoders and then projected and resampled into the LLM's token embedding space. All media tokens are then interleaved in the query template defined in Eq. (1). Finally, we produce a single embedding vector by aggregating the LLM's output hidden states. The same model is used to encode both the query and the candidate. We keep the media encoders and the LLM frozen, training only our novel modules and a LoRA [20] adapter injected into the LLM. Our novel components are detailed below.

Shared Media Resampler. A key challenge in multimodal systems is the high token count produced by media (visual and audio) encoders (often $>$ 500 tokens), which limits training batch sizes and computational efficiency. To address this, we introduce a Shared Media Resampler (Fig. 3, left). This module is an intermediate layer connecting media tokens with the LLM input space. It condenses a large sequence of media tokens into a small, fixed number of latent vectors using the Perceiver [23] architecture.

We customize this module to leverage data from different modalities. While a single perceiver module is shared across all media types, we introduce modality-specific latents. These are added to the shared latent queries, enabling the module to adapt to each modality type while retaining its generalization capability. We use all media tokens from the media encoders as input for the resampler, except for videos, where we first apply 3D trilinear interpolation to the video features to reduce frame-level redundancy before resampling.

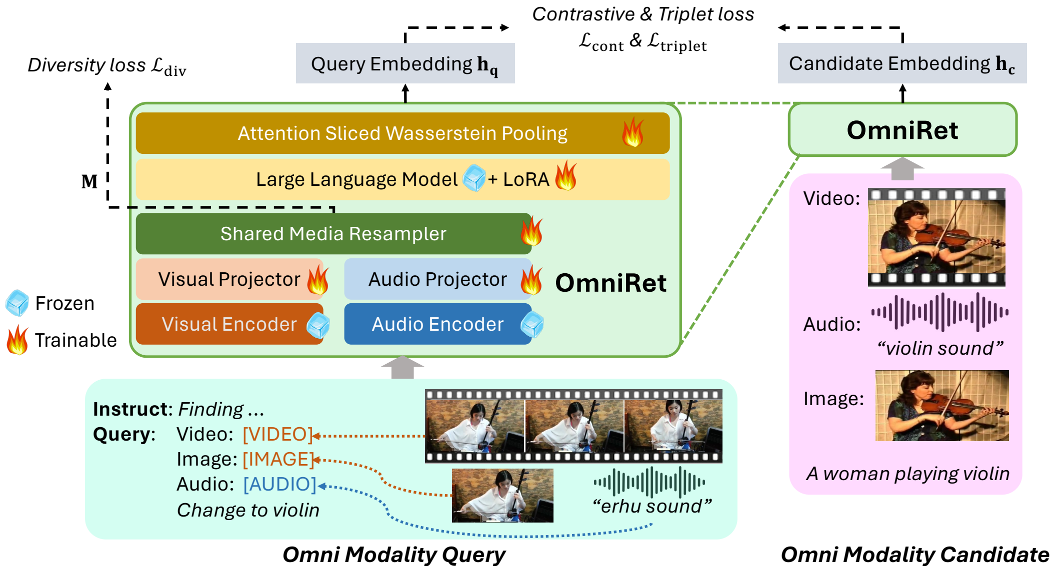


Figure 2. **Overall OmniRet architecture.** Our universal retrieval model integrates specialized visual and audio encoders with a Large Language Model acting as a cross-modal composer. Media inputs are processed through a Shared Media Resampler, and the final embedding is derived via our Attention Sliced Wasserstein Pooling. The model is optimized using Contrastive, Triplet and Diversity losses.

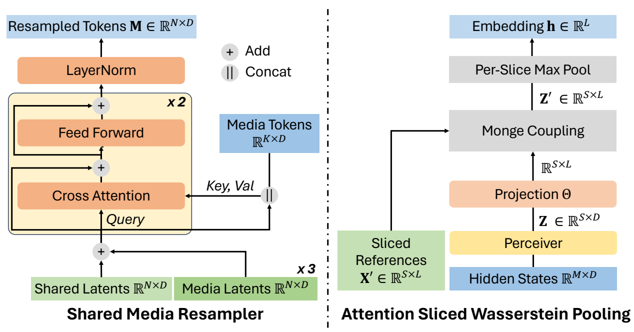


Figure 3. *Left:* Our **Shared Media Resampler** condenses the output of media encoders into a compact set of latent vectors before they are fed to the LLM. *Right:* Our **Attention Sliced Wasserstein Pooling (ASWP)** aggregates the final LLM output hidden states into a single, high-fidelity embedding vector. Attention Resampler applies to both output of media encoders and LLM universal encoder.

Diversity on Media Resampled Tokens. To ensure the resampled tokens capture diverse information, we apply a diversity regularization loss, $\mathcal{L}_{\text{div}}$, which encourages orthogonality among the output vectors $\mathbf{M} \in \mathbb{R}^{N \times D}$:

$$\mathcal{L}_{\text{div}} = \frac{1}{N^2} \sum \text{smooth}_{L1} \left(\text{Dropout} \left(\max(\mathbf{M}\mathbf{M}^\top, \mathbf{0}) - \mathbf{I} \right) \right) \tag{5}$$

Here, we compute the pairwise similarity matrix $\mathbf{M}\mathbf{M}^\top$, clip negative values, and then remove the self-similarity scores (the identity matrix $\mathbf{I}$). Crucially, we apply Dropout to the resulting matrix before the loss. This acts as a sparse sampling mechanism, ensuring that the loss is computed on only a small, randomly selected subset of token pairs in each training step. This efficiently encourages global diversity.

The smooth$_{L1}$ function (with $\gamma = 0.5$) is then applied. This Huber loss is less sensitive to large similarities (outliers) than a standard $L2$ loss, preventing exploding gradients while still penalizing non-orthogonality. The smooth$_{L1}$ definition is as follows:

$$\text{smooth}_{L1}(x) = \begin{cases} 0.5x^2/\gamma, & \text{if } x < \gamma \\ x - 0.5 * \gamma, & \text{otherwise} \end{cases} \tag{6}$$

Attention Sliced Wasserstein Pooling. To aggregate the LLM output hidden states, we first compress the full sequence into a compact set of S latent embeddings, $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_S \mid \mathbf{z}_i \in \mathbb{R}^D\}$, using an attention-based resampler identical to the Shared Media Resampler in Fig. 3 (left), applied to the LLM tokens. Instead of average pooling $\mathbf{Z}$, which can blur fine-grained token structure, we use Attention Sliced Wasserstein Pooling (ASWP) to summarize the distribution of $\mathbf{Z}$ relative to a set of learnable reference points $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_S \mid \mathbf{x}_i \in \mathbb{R}^D\}$.

Inspired by PSWE [52], ASWP compares $\mathbf{Z}$ and $\mathbf{X}$ across L 1D projections $\Theta = \{\theta_1, \ldots, \theta_L \mid \theta_i \in \mathbb{S}^{D-1}\}$. For each projection θ_i, we project both sets onto θ_i and compute their 1D Monge coupling, producing

$$\mathbf{Z}' = [\psi_1(\mathbf{X}, \mathbf{Z}; \theta_1); \ldots; \psi_L(\mathbf{X}, \mathbf{Z}; \theta_L)] \in \mathbb{R}^{S \times L}, \tag{7}$$

where $\psi_i(\cdot) \in \mathbb{R}^S$ measures how the projected token distribution aligns with the projected references. Thus, $\mathbf{Z}'$ can be viewed as a learnable histogram-like descriptor that preserves fine-grained distributional information.

Since $\mathbf{Z}'$ is larger than the target embedding size, ASWP further aggregates it with a hard selection step. For each col-

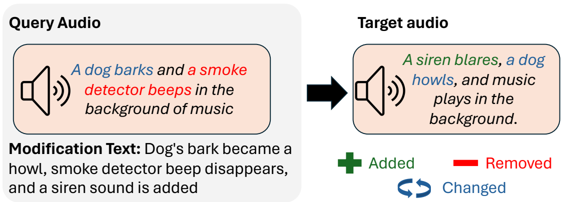


Figure 4. A composed audio retrieval example from our ACM benchmark where query composes both audio and text modalities.

umn ψ_i of $\mathbf{Z}'$, we compute a soft score $\mathbf{y} = \text{softmax}(\psi_i) \in [0,1]^S$ and form a one-hot mask

$$m_i^{\text{hard}} = \text{OneHot}\left(\arg\max_{0\leq j<S} \mathbf{y}_j\right). \tag{8}$$

To backpropagate through this discrete choice, we use the straight-through maximum (STM) estimator:

$$\tilde{m}_i = m_i^{\text{hard}} - \text{StopGrad}(\mathbf{y}) + \mathbf{y}. \tag{9}$$

In the forward pass, $\tilde{m}_i$ is equivalent to the binary mask m_i^{hard}, while in the backward pass it allows gradients to flow through $\mathbf{y}$. We then apply this selection by Hadamard product, $\mathbf{V} = \mathbf{Z}' \odot \tilde{m}(\mathbf{Z}')$, and obtain the final L-dimensional embedding $\mathbf{h}$ by column-wise summation: $\mathbf{h}_i = \sum_{j=1}^{S} \mathbf{V}_{ji}$. This procedure is illustrated in Fig. 3 (right). In practice, we learn the sliced references of $\mathbf{X}'$ directly.

Final Loss. The final loss to train OmniRet is the linear combination of $\mathcal{L}_{\text{cont}}, \mathcal{L}_{\text{triplet}}$ and $\mathcal{L}_{\text{div}}$:

$$\mathcal{L} = \mathcal{L}_{\text{cont}} + \mu_1 \mathcal{L}_{\text{triplet}} + \mu_2 \mathcal{L}_{\text{div}} \tag{10}$$

where $\mu_1 = 1$ and $\mu_2 = 0.1$ is set in our experiments.

4. Audio-Centric Multimodal Benchmark

While composed image retrieval has been extensively investigated [47, 59, 73], no curated benchmark exists for the composed audio retrieval task. Establishing such a benchmark is essential to advance research in complex audio retrieval. Furthermore, few benchmarks exist for audio-to-image or audio-to-video retrieval, largely due to the limited capacity of most models to process such heterogeneous query modalities. To address these gaps, we introduce the Audio-Centric Multimodal (ACM) benchmark, which comprises two new tasks: composed audio retrieval (audio-text to audio) and audio-visual retrieval (audio to image/video).

4.1. Data collection

Audio Source. We use the popular VGG-Sound dataset [6], which is widely used for audio-visual tasks. This choice is based on two rationales. First, it contains a large volume of audio-visual correspondence data in the wild, making it well-suited for our audio-visual retrieval task. Second, the composed audio retrieval task requires query-target audio pairs that are sufficiently similar yet exhibit clear distinctions. VGG-Sound, which comprises 200k 10-second audio clips from ∼310 classes, satisfies this requirement.

Table 1. **Subjective evaluation of our ACM benchmark.** For the "Gemini-2.5 w/ text only" method, the audio inputs were replaced with their corresponding generated captions.

Evaluation Method	Accuracy
Human	87%
Gemini-2.5 w/ text only	96%

Representative Audio Selection. We leverage the ground-truth class labels from VGG-Sound to randomly select 30 audio-video pairs per class. In total, we sample 9,270 pairs to form a representative subset, denoted as $\mathcal{D}$. From this set, we first construct the query-target pairs for the composed audio task, and then reuse these samples for the audio-visual retrieval task.

Query-Target Filtering. Our filtering process aims to find a set of 3-5 target audios for each query audio $A \in \mathcal{D}$ that are semantically similar but not identical. Since VGG-Sound lacks captions, we first generate a caption for every audio in VGG-Sound using QwenOmni2.5 [76]. Let κ be the Sentence-BERT [60] cosine similarity between the caption of A and the caption of another audio in the dataset. To find targets for A, we first rank the top 50 audios by κ in descending order. From this ranked list, we select up to 5 audios that satisfy three criteria: *(1) Similarity:* The caption similarity must be in the range $0.6 < \kappa < 0.8$, ensuring both relevance and diversity. *(2) Textual Diversity:* A candidate is excluded if its κ differs by less than 0.05 from the most recently added target, thereby maintaining sufficient distinctiveness among selected samples. *(3) Acoustic Diversity:* A similar criterion to (2) is applied, but using CLAP [74] audio embeddings, ensuring distinctiveness in the acoustic information. We terminate the selection process for A once 5 targets are found and discard A entirely if fewer than 3 targets can be found. After this filtering process, we retain 1,292 audios from $\mathcal{D}$ as our final query set. Since each of these 1,292 query audios has 3 to 5 associated targets, we obtain a total of 4,251 target audios.

Benchmark Construction. For each (query, target) audio pair identified, we generate a modification sentence describing the difference between them. We employ the Gemini2.5-Flash-Lite model [9] in "thinking mode", prompting it with the captions of the query-target pair (details in supplementary). This process constructs the composed audio benchmark as a set of triplets: (query-audio, modification-text, target-audio). Fig. 4 shows an example.

Furthermore, we construct the audio-visual retrieval task by reusing the 1,292 audio queries. The benchmark consists of (audio, video) and (audio, image) pairs. The target

Table 2. **Performance comparison of OmniRet with state-of-the-art baselines across 13 retrieval tasks on the extended version of M-BEIR benchmark.** Our model demonstrates strong, balanced performance on established visual-text tasks while uniquely extending capabilities to the audio modality. Modalities: I (Image), V (Video), A (Audio), T (Text). **Bold** and **green** indicate the best performance in each group and overall, respectively. MMEmbed [40] is included for reference only, as it uses a larger LLM and is not directly comparable.

Model	LLM Size	Uni-Modality		Textual Binding						Composed Visual Text				
		I→I	T→T	I→T	T→I	V→T	T→V	A→T	T→A	T→I,T	I,T→T	I,T→I	I,T→I,T	V,T→V
Text-Binding Pretrained Models														
CLIP [57]	-	25.9	40.5	43.1	37.5	14.2	15.9	-	-	-	-	-	-	-
SigLIP [78]	-	25.9	34.0	56.5	52.1	19.7	22.5	-	-	-	-	-	-	-
PE-Core [3]	-	**32.0**	**56.8**	**58.0**	**53.4**	**32.3**	**29.5**	-	-	-	-	-	-	-
MMT [54]	-	-	-	-	-	-	-	49.8	46.8	-	-	-	-	-
CLAP [74]	-	-	-	-	-	-	-	**63.9**	**56.6**	-	-	-	-	-
Multi-Task Finetuning Models														
CLIP$_{SF}$ [71]	-	28.4	83.7	50.0	45.1	-	-	-	-	63.6	41.0	26.4	60.6	-
MMEmbed [40]	7B	32.1	96.7	49.9	44.2	-	-	-	-	77.2	47.2	38.5	68.5	-
VLM2VecV2 [50]	1.5B	**30.0**	81.1	43.4	39.8	17.6	18.4	-	-	61.6	24.5	28.7	33.6	76.4
OmniRet	1.5B	24.4	**86.7**	**50.6**	**46.9**	**43.8**	**43.2**	**66.8**	**62.4**	**70.5**	**44.4**	**36.5**	**64.8**	**86.2**

videos are from the original 1,292 query audio-video samples, and the target images are extracted from their middle frames. The remaining audio-video samples from $\mathcal{D}$ (that were not selected as queries) are added to the candidate pool, resulting in 1,292 queries and 5,480 total candidates.

4.2. Data Quality Analysis

We conducted a subjective evaluation to verify the quality of our generated captions and modification texts. We tasked three human annotators with answering 100 multiple-choice questions, each created from a randomly sampled query audio. Each question presented the query audio and the modification text; the answer choices consisted of the correct target audio and two random distractor audios from the same candidate subset. As shown in Table 1, human annotators achieved 87% accuracy, indicating the task is solvable but non-trivial. In contrast, the Gemini-2.5 model (using text-only inputs) achieved 96% accuracy. These results confirm that our generated audio captions and modification texts are high-quality and sufficiently descriptive for the composed audio retrieval task.

5. Experiments

5.1. Implementation Details

Training Datasets. We extend M-BEIR [71] with additional text, image, video, and audio retrieval datasets. Specifically, we added 7 datasets (MSMarco [2], HotpotQA [77], NaturalQuestion [31], PAQ [35], StackExchange [62], NLI [4], SQuAD [58]) from MTEB [64] for text retrieval following MMEmbed [41]; 3 datasets (LLaVA-558K [44], CC-CoIR [67], MTCIR [22]) to enhance the interaction between image and text; 5 video and text datasets (TGIF [38], Charades [61], WebVid2M [1], PE-Video [3], WebCoVR [67]); 4 datasets for audio and text retrieval (AudioCaps [29], ClothoV2.1 [11], WavText5K [10], WavCaps [49]) and VGGSound [6] for audio-visual retrieval. In total, we use 30 datasets for training, totaling approximately 6.2 million query-candidate pairs. The distribution and details of the training datasets can be found in the supplementary material.

Benchmarks and Evaluation Metrics. Besides our proposed ACM benchmark, we also evaluate models on an extended version of M-BEIR and a subset of the benchmarks from MMEBv2 [50]. For M-BEIR, we extended it to include additional interactions between the audio-text and video-text modalities. Following MMEBv2, we test each benchmark individually and average the results for each group instead of using a global candidate pool. We use Recall@k with $k = 5$ for most benchmarks, except for FashionIQ [73] and Fashion200K [18], which use $k = 10$. On MMEBv2, we compare performance on classification (CLS), retrieval (RET), and video moment retrieval tasks (MRET). Recall@1 is used as the metric for MMEBv2, while on our proposed benchmark, we report Recall@5. We follow the evaluation protocols of prior works for all models and baselines.

Baselines. On the extended M-BEIR, we evaluate against dual-encoder models trained on text-binding tasks (*e.g.*, image-to-text) and recent baselines finetuned on multi-modal (image and video) data. For MMEBv2 baselines, we gather results for models with LLM backbones smaller than 4B parameters to be comparable with our model. On our benchmark with the audio modality, we choose to compare with ImageBind [16], CLAP [74], ViT-Lens [34], OmniBind [70] and Gemma Embedding [68]. While ImageBind, ViT-Lens, and OmniBind are strong baselines for audio-image-video alignment, the Gemma encoding captions from QwenOmni [76] serves as an alternative solution for the composed audio retrieval task. Details of each baseline are discussed in the supplementary material.

Model Architecture and Training Details. At the core of our architecture, we employ GTE-Qwen2-1.5B-Instruct [39] as the universal LLM for processing and

Table 3. **Generalization performance of OmniRet against leading models (<7B parameters) on a subset of MMEBv2.** Our model achieves outstanding performance on video tasks. Notably, it maintains competitive performance on image retrieval tasks without being fully fine-tuned on their training sets. **Bold** highlights the best score in each group.

Model	LLM Size	Image		Video		
		CLS	RET	CLS	RET	MRET
Image-Text Embedding Models						
UniME [17]	1.5B	59.0	64.9	-	-	-
LLaVE [32]	1.5B	62.1	65.2	-	-	-
B3++ [65]	1.5B	67.0	70.9	-	-	-
MetaEmbed [75]	2B	**68.1**	**71.9**	-	-	-
Video-Image-Text Embedding Models						
ColPaliv1.3 [14]	2B	40.3	48.1	26.7	21.6	25.5
GME [81]	1.5B	54.4	66.9	34.9	25.6	32.4
VLM2VecV1 [27]	1.5B	58.7	65.0	33.4	20.6	33.0
VLM2VecV2 [50]	1.5B	**62.9**	**69.5**	39.3	28.8	38.5
OmniRet	1.5B	51.7	65.3	**48.6**	**36.5**	**43.3**

fusing omnimodal information into a coherent embedding space. Each media input is processed by a different encoder. While visual inputs are handled by SigLIP-SO400M-Patch14-384 [78], audio files are processed by QwenAudio Encoder [8]. Our entire implementation is built upon the publicly available LLaVA [44] codebase. To balance task performance and maintain a large number of in-batch hard negatives, we customized a data sampler that balances the number of samples between tasks within the same batch and randomly selecting one or two datasets per task. With the majority of parameters remaining frozen, our model has only $\sim$84M trainable parameters in total.

Our training protocol is divided into two distinct stages:

- **Stage 1: Warm-up.** We first warm up the model on a curriculum of simple uni-modality and textual binding retrieval tasks, excluding video-related and composed-query datasets. Only the projectors, resampler, and pooling layer are trained while the LLM is kept frozen. We train the model on 2M samples with the batch size of 2048, approximately 341 in-batch samples per task.
- **Stage 2: Fine-tuning.** We continue to train the model on all datasets and tasks for approximately 18M training samples. The video media latents are initialized from the weights of image media latents. Besides trainable modules from stage 1, LoRA with a rank of 16 and alpha of 64 is applied to the LLM. We train with a batch size of 3072, but select only 4 random tasks per batch and apply gradient accumulation in 2 steps to keep training stable and robust across all tasks.

More details on the training hyper-parameters can be found in the supplementary material.

5.2. Quantitative Results

Table 2 compares our OmniRet with other methods, both with and without training on the M-BEIR datasets. The first group includes methods trained only on textual binding

Table 4. **Performance comparison on our novel ACM benchmark.** OmniRet illustrates promising results over baselines on both the composed audio and audio-visual retrieval tasks. Modalities: I (Image), V (Video), A (Audio), T (Text). **Bold** and underline indicate the best and second-best performance, respectively.

Model	A,T→A	A→V	V→A	A→I	I→A
QwenOmni [76] + Gemma [68]	**44.6**	3.3	6.3	4.4	5.4
CLAP [74]	16.1	-	-	-	-
ImageBind [16]	7.32	**35.5**	**36.3**	**30.1**	**29.7**
ViT-Lens [34]	10.1	-	-	24.7	25.9
OmniBind [70]	15.6	-	-	30.5	29.1
OmniRet	23.0	**35.5**	34.4	24.5	26.0

tasks; while they show strong performance on those specific tasks, their ability is limited as they cannot handle complex data structures or different modality types. For visual-text retrieval, PE-Core [3] is presented as the strongest baseline, with leading performance on most tasks. For audio, we compare our model with two strong baselines, CLAP [74] and MMT [54]. Among models fine-tuned on multi-task retrieval, CLIP$_{SF}$ [71] and MMEmbed [41] are fine-tuned on the M-BEIR dataset. VLM2VecV2 [50] is another strong baseline, which is fine-tuned on a broader set of datasets that includes the video modality. In comparison with models of a similar size, OmniRet achieves leading performance on all tasks except for image-to-image retrieval, often by a large margin. Notably, on audio and video tasks, OmniRet also outperforms all specialized models, even those trained on larger amounts of in-domain data.

We then evaluate OmniRet's generalization on a subset of tasks from MMEBv2, shown in Table 3. Since our model was not fine-tuned on the training sets for the Visual Question Answering or Visual Document Retrieval tasks, we only compare performance on classification (CLS) and retrieval (RET) tasks. Our model maintains highly competitive performance on image retrieval, achieving a score (approximately 65) that is within the median range of models fully fine-tuned on this data. For video, OmniRet achieves state-of-the-art performance even without being fully trained on all datasets, demonstrating the strong generalization ability of our framework.

Finally, on our new ACM benchmark (Table 4), we compare OmniRet with five baselines. For composed audio retrieval (A,T→A), the strongest baseline is a multi-stage text-to-text retrieval pipeline (using Gemma [68] with QwenOmni2.5 [76]-generated captions). Although it shows promising results, this multi-stage pipeline incurs significant additional computational cost. On this same task, we also show the performance of CLAP [74], ImageBind [16], ViT-Lens [34], and OmniBind [70]with features fused as UniIR [71]. OmniRet outperforms by a large margin while remaining an efficient single-stage method. The Gemma-based pipeline, while strong on the composed audio task, fails to work well on other tasks in our benchmark. This is likely because converting audio and images/videos to tex-

Table 5. **Ablation study of our proposed components.** We report the impact on Average Recall (across 6 tasks, trained on 1M samples) when each component is removed or modified.

	Avg. Recall	Δ
Our Baseline	50.2	0.0
(a) Embedding Vector		
Single [EOS] Vector	43.4	-6.8
Multi (16) Vectors	49.8	-0.4
(b) Number of Projections (L) and References (S)		
$L = 1536, S = 128$	47.5	-2.7
$L = 1024, S = 128$	48.7	-1.5
$L = 4096, S = 64$	48.7	-1.5
$L = 4096, S = 32$	48.1	-2.1
(c) Pooling in ASWP		
Average Pooling	20.7	-29.5
Learnable Weighted Sum	48.5	-1.7
Maximum Pooling	49.2	-1.0
(d) Media Resampler		
No Resampler	46.7	-3.5
Separated Resamplers	49.9	-0.3
Shared Resampler W/o Media Latents	49.8	-0.4
(e) Loss Functions		
W/o $\mathcal{L}_{\text{triplet}}$	49.7	-0.5
W/o $\mathcal{L}_{\text{div}}$	47.1	-3.1

tual descriptions breaks the modality binding between audio and vision, thereby substantially degrading audio-visual retrieval performance. In contrast, ImageBind, ViT-Lens, and OmniBind have the capacity to align visual and audio embeddings, achieving strong performance on all audio-visual retrieval tasks. However, these models lack the capacity to understand the complex, composed queries in the A,T→A task. Our model, OmniRet, demonstrates a promising ability to align both visual and audio modalities, achieving performance similar to ImageBind on audio-to-video retrieval and remaining competitive on other audio-visual tasks, all while dominating the complex composed retrieval task.

5.3. Ablation Studies

We conduct comprehensive ablation studies, training each configuration for 1M samples and reporting the Average Recall across 6 tasks in Table 5. For resource-intensive settings (no resampler and multi-vectors), we halved the batch size and doubled the steps (with gradient accumulation) for a fair comparison. Our default model uses a Shared Media Resampler, ASWP pooling ($L = 4096, S = 128$) with the Straight-Through Maximum (STM) estimator, and all three loss functions ($\mathcal{L}_{\text{cont}}, \mathcal{L}_{\text{triplet}}, \mathcal{L}_{\text{div}}$).

(a) ASWP Improves Embedding Quality. Using a single '[EOS]' vector as the embedding cuts performance by 6.8%. Late interaction (16 vectors) is better than '[EOS]' but slightly worse than ASWP, likely due to the required smaller batch size limiting hard-negative sampling.

(b) More Projectors and References are Beneficial. Increasing projections L and references S improves results at the cost of computation. We find $L = 4096, S = 128$ offers the best tradeoff.

(c) STM is the Best for ASWP. Average pooling performs poorly, as it cancels out informative positive distances to the references with other negative values. Our STM estimator is the best, outperforming standard max pooling and a learnable weighted sum by enabling better gradient flow.

(d) Shared Media Resampler is the Best Choice. Removing the resampler degrades performance (due to a smaller batch size). Adding our resampler brings a +3% recall boost. Using separate resamplers or a shared one without our modality-specific latents both perform worse, showing that sharing data and maintaining specificity is key.

(e) Triplet and Diversity Losses are Essential. The result is slightly degraded (0.5% drop) when removing the $\mathcal{L}_{\text{triplet}}$ loss. In contrast, performance drops significantly (3.1% drop) without the $\mathcal{L}_{\text{div}}$ loss. This highlights the critical importance of maintaining diversity on the resampled media tokens before they are fed to the LLM.

6. Limitations and Conclusion

Despite promising results, OmniRet has clear limitations that open avenues for future work. First, due to resource constraints, we leave scaling the model with a larger LLM backbone and more training data for future investigation, which we expect would significantly boost performance. Second, our work focuses on expanding modalities, whereas other works (*e.g.*, MMEBv2) have expanded retrieval tasks. A key future direction is to train a single, unified model on a large-scale dataset incorporating both more tasks and more diverse input types (*e.g.*, depth maps, 3D point clouds, and speech). Finally, while we introduce a new benchmark for audio retrieval, it could be extended to more complex scenarios, such as retrieval over interleaved, mixed-media documents.

In conclusion, we present a novel approach to omni-modality retrieval where image, video, audio, and text are encoded into a unified embedding space by a single model. Our contributions include:(1) An innovative architecture that encodes any modality into a shared embedding space, considering both the diversity and efficiency of the extracted information; (2) A new Sliced Wasserstein embedding pooling method that overcomes the limitations of coarse–grained, single-vector embeddings; (3) A new training strategy and dataset combination that improve the model's generalization across many retrieval tasks; (4) New ACM benchmark for evaluating interactions between audio and other modalities. Notably, OmniRet achieves superior results with 12/13 leading performance on extended M-BEIR, outstanding results on video benchmarks on MMEBv2 and promising performance on ACM. These advancements collectively push the boundaries of omni-modality retrieval, offering more efficient and effective solutions for real-world applications like RAG or recommendation systems.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 6, 5

[2] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016. 6, 5

[3] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Abdul Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. In *NeurIPS*, 2025. 6, 7, 4, 5, 8

[4] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015. 6, 5

[5] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *CVPR*, 2022. 5

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*. IEEE, 2020. 5, 6

[7] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*, 2023. 5

[8] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023. 7

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2, 5

[10] Soham Deshmukh, Benjamin Elizalde, and Huaming Wang. Audio retrieval with wavtext5k and clap training. *arXiv preprint arXiv:2209.14275*, 2022. 6, 5

[11] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP*. IEEE, 2020. 6, 5

[12] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP*. IEEE, 2023. 2, 3

[13] Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in CLIP. In *ICLR*, 2025. 2

[14] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024. 7, 4

[15] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 5

[16] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 3, 6, 7

[17] Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. Breaking the modality barrier: Universal embedding learning with multimodal llms. In *ACM MM*, 2025. 2, 7, 4

[18] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *ICCV*, 2017. 6, 5

[19] D Hendrycks. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 1

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 3

[21] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *ICCV*, 2023. 5

[22] Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, Trishul Chilimbi, and Abhinav Shrivastava. Collm: A large language model for composed image retrieval. In *CVPR*, 2025. 3, 6, 5

[23] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, 2021. 3

[24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*. PMLR, 2021. 2

[25] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024. 2

[26] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024. 2, 4

[27] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2vec: Training vision-language models for massive multimodal embedding tasks. In *ICLR*, 2025. 2, 7, 4

[28] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020. 2, 3

[29] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019. 6, 5

[30] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *NeurIPS*, 2022. 4

[31] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *ACL*, 2019. 6, 5

[32] Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. Llave: Large language and vision embedding models with hardness-weighted contrastive learning. *arXiv preprint arXiv:2503.04812*, 2025. 7, 4

[33] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-embed: Improved techniques for training LLMs as generalist embedding models. In *ICLR*, 2025. 3

[34] Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, Dylan Sun, Yuying Ge, Ying Shan, and Mike Zheng Shou. Vit-lens: Towards omni-modal representations. In *CVPR*, 2024. 6, 7, 4

[35] Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. Paq: 65 million probably-asked questions and what you can do with them. *ACL*, 2021. 6, 5

[36] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 2022. 1, 2

[37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*. PMLR, 2023. 1, 2

[38] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *CVPR*, 2016. 6, 5

[39] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023. 6

[40] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. In *ICLR*, 2024. 6, 8

[41] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. MM-EMBED: UNIVERSAL MULTIMODAL RETRIEVAL WITH MULTIMODAL LLMS. In *ICLR*, 2025. 2, 6, 7, 4

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[43] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743*, 2020. 5

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 6, 7, 5

[45] Siqi Liu, Weixi Feng, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. Edis: Entity-driven image search over multimodal web content. *arXiv preprint arXiv:2305.13631*, 2023. 5

[46] Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *CVPR*, 2025. 2

[47] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, 2021. 5

[48] Manh Luong, Khai Nguyen, Nhat Ho, Gholamreza Haffari, Dinh Phung, and Lizhen Qu. Revisiting deep audio-text retrieval through the lens of transportation. In *ICLR*, 2024. 2

[49] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024. 6, 5

[50] Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025. 2, 6, 7, 8

[51] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 4

[52] Navid Naderializadeh, Joseph F Comer, Reed Andrews, Heiko Hoffmann, and Soheil Kolouri. Pooling by sliced-wasserstein embedding. *NeurIPS*, 2021. 2, 3, 4

[53] Khai Nguyen. An introduction to sliced optimal transport: foundations, advances, extensions, and applications. *Foundations and Trends® in Computer Graphics and Vision*, 2025. 2

[54] Andreea-Maria Oncescu, A Koepke, Joao F Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries. *arXiv preprint arXiv:2105.02192*, 2021. 6, 7, 4, 8

[55] Khoi Pham, Chuong Huynh, Ser-Nam Lim, and Abhinav Shrivastava. Composing object relations and attributes for image-text matching. In *CVPR*, 2024. 3

[56] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In *CVPR*, 2023. 3

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PmLR, 2021. 1, 2, 6, 4, 8

[58] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. 6, 5

[59] Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan Plummer, Ranjay Krishna, and Kate Saenko. Cola: A benchmark for compositional text-to-image retrieval. *NeurIPS*, 2023. 5

[60] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*. Association for Computational Linguistics, 2019. 5

[61] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018. 6, 5

[62] Stack-Exchange-Community. Stack exchange data dump. *ACL*, 2023. 6, 5

[63] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *ArXiv*, abs/2402.04252, 2024. 2

[64] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021. 6

[65] Raghuveer Thirukovalluru, Rui Meng, Ye Liu, Mingyi Su, Ping Nie, Semih Yavuz, Yingbo Zhou, Wenhu Chen, Bhuwan Dhingra, et al. Breaking the batch barrier (b3) of contrastive learning via smart batch mining. *arXiv preprint arXiv:2505.11293*, 2025. 7, 4

[66] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2, 4

[67] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr-2: Automatic data construction for composed video retrieval. *IEEE TPAMI*, 2024. 6, 5

[68] Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025. 6, 7, 2

[69] Ziyang Wang, Heba Elfardy, Markus Dreyer, Kevin Small, and Mohit Bansal. Unified embeddings for multimodal retrieval via frozen llms. In *Findings of the Association for Computational Linguistics: EACL 2024*, 2024. 2

[70] Zehan Wang, Ziang Zhang, Minjie Hong, Hang Zhang, Luping Liu, Rongjie Huang, Xize Cheng, Shengpeng Ji, Tao Jin, Hengshuang Zhao, et al. Omnibind: Large-scale omni multimodal representation via binding spaces. In *ICLR*, 2025. 6, 7, 4

[71] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *ECCV*. Springer, 2024. 2, 6, 7, 1, 3, 4, 8

[72] Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. *arXiv preprint arXiv:2508.21038*, 2025. 2, 3

[73] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *CVPR*, 2021. 5, 6

[74] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*. IEEE, 2023. 1, 5, 6, 7, 3, 4, 8

[75] Zilin Xiao, Qi Ma, Mengting Gu, Chun-cheng Jason Chen, Xintao Chen, Vicente Ordonez, and Vijai Mohan. Metaembed: Scaling multimodal retrieval at test-time with flexible late interaction. *arXiv preprint arXiv:2509.18095*, 2025. 3, 7, 4

[76] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 2, 5, 6, 7

[77] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018. 6, 5

[78] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 2, 6, 7, 8

[79] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: self-supervised image retrieval with open-ended instructions. In *International Conference on Machine Learning*, 2024. 3

[80] Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *ICLR*, 2025. 2

[81] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024. 2, 7, 5

[82] Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. MegaPairs: Massive data synthesis for universal multimodal retrieval. In *ACL*, pages 19076–19095, Vienna, Austria, 2025. Association for Computational Linguistics. 2, 3