Omni-Interactive Universal Embedder

Wei-Yao Wang[*], **Kazuya Tateishi**, **Shuyang Cui**, **Christian Simon**,
Takashi Shibuya, **Shusuke Takahashi**, **Yuki Mitsufuji**,
Sony Group Corporation, Sony AI

Abstract

Multimodal representation learning has been shifting from traditional two-tower architectures to large language model (LLM)-based embedders due to their strong instruction-following capabilities. Despite this progress, existing approaches primarily focus on language and image modalities, which also remain the dominant modalities for user-conditioned interactions in current embedders. In this paper, we propose the first Omni-Interactive Universal Embedder (OmniUE), which not only learns a unified embedding space across text, video, and audio by leveraging intermediate-layer representations from dedicated learnable tokens, but also supports omni-interactive querying, enabling users to provide inputs in the form of text, visual regions of interest, and audio spans. Within OmniUE, visual and audio segmenters process diverse user interactions and integrate them with an omni-LLM to produce user-conditioned any-to-any embeddings via context aggregation. To evaluate OmniUE's omni-interactive capabilities, we introduce OmniCHOIR, benchmarking models for omni-interactive compositional audio retrieval based on the given text, video, and audio as well as unimodal or multimodal interaction prompts. OmniUE consistently surpasses state-of-the-art baselines across diverse modalities, with average improvements of 10.4% on textual-interactive video benchmarks (MMEB-v2-video), 1.0% on audio tasks (MAEB), 83.1% on visual-interactive benchmarks (SCaR), and 21.9% on our omni-interactive OmniCHOIR benchmark. We believe that jointly advancing omni-modal representation learning and omni-interactive querying paves the way toward universal embedders. Project page: https://wywywang.github.io/omniue-demo/.

1 Introduction

Multimodal embedding models have advanced from traditional two-tower-based models (e.g., [47, 31, 67, 55]) to multimodal LLM (MLLM)-based models due to their large-scale training corpora, support for arbitrary combinations of input modalities, and effective instruction-following abilities that allow users to extract corresponding embeddings from the same embedder using different text instructions. This transition has been demonstrated to be effective in terms of reasoning and understanding over images [27, 35, 69, 34, 8, 56] and video [38, 22, 29]; however, they only focus on text with the visual modality. Recently, researchers have started exploring omni-modal embedders by finetuning omni-LLMs to encode heterogeneous inputs to a unified embedding space [9, 58, 64, 36], which opens the way toward universal embedders, i.e., general-purpose embedders that support diverse modalities, tasks, and interaction scenarios within a unified framework.

Despite this progress toward universal embedders, current MLLM-based approaches mainly focus on incorporating new modalities with text or improving text-based instruction following, with text remaining the dominant medium for user interaction. Since text alone is often insufficient to precisely express complex or fine-grained user intent, Wang et al. [53] enable visual prompts as an additional

[*]Project Lead: Wei-yao.Wang@sony.com.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

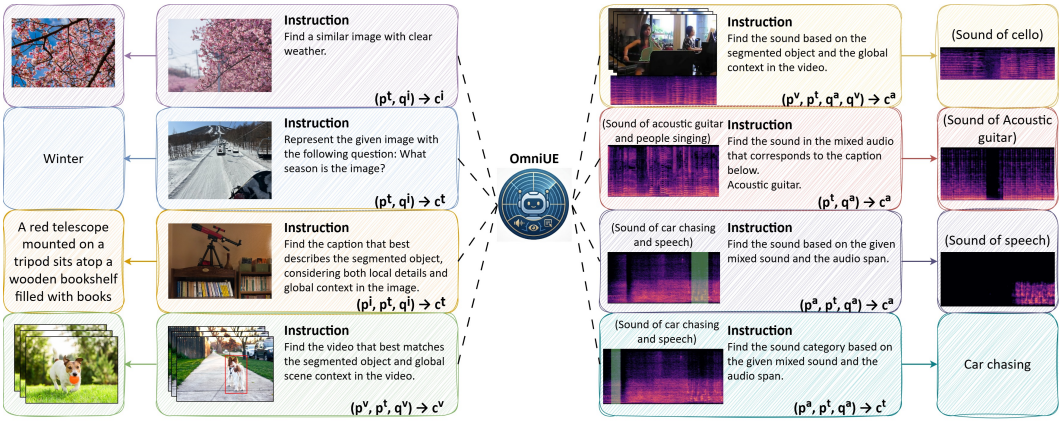

Figure 1: Applicability of OmniUE, which takes omni-interactive prompts (p) and queries (q) to retrieve any-modality targets (c), including text, image, video, and audio.

human-machine interaction modality to embedders, yet their approach still operates within the image-text paradigm. We argue that humans should be able to interact with embedding models not only through textual and visual guidance, but also via video and audio modalities as interaction modalities, which we term *omni-interactions*. For instance, visual prompts may fall short in video scenarios involving both on-screen and off-screen sources or ambiguous audio-visual correspondences, where selecting a spatial region alone cannot reliably capture the intended semantics. In addition, in complex auditory scenes with overlapping or intermittent sound sources (e.g., retrieving segments that contain periodic bird chirps embedded in background noise, rather than all clips matching the coarse label "bird"), textual queries are often insufficiently precise to capture retrieval intent, while visual cues may be absent or weakly correlated, necessitating direct audio-based interaction. Enabling omni-interactions in embedding models not only facilitates more accurate and user-aligned representations, but also advances toward truly universal embedders for user-conditioned interactions, with applications beyond retrieval, including context engineering [45], agentic systems [40, 71, 59], LLM memory augmentation [42, 15], and personalization [32]. These challenges give rise to two central yet underexplored questions: *1) How can we enable omni-interactions in omni-modal embedding models? 2) Given such capabilities, how can we systematically evaluate individual or joint interactions within multimodal embedding benchmarks?*

In this paper, we propose OmniUE, a novel omni-interactive universal embedder that supports optional omni-modal interactions and arbitrary combinations of text, silent video, and audio inputs to produce a unified representation. Specifically, OmniUE is built upon two complementary axes: i) a query stream, which encodes multimodal inputs (visual, textual, and audio) into holistic contextual representations, and (ii) an interactive prompt stream, which captures user-specified signals as entity- or intent-level guidance. To learn multi-level context across complex multi-modalities, we introduce additional learnable tokens appended in the sequence and aggregate them from intermediate LLM layers into a final single embedding. These axes enable flexible and fine-grained conditioning of embeddings through heterogeneous interaction signals, while maintaining a unified representation space. To the best of our knowledge, OmniUE is the first framework to generate user-conditioned unified embeddings across arbitrary modality combinations (as shown in Fig. 1), generalizing beyond prior approaches that rely primarily on textual instructions[2]. Importantly, such textual interaction can be viewed as a special case within our broader omni-interaction paradigm, highlighting the generality and extensibility of our formulation.

Since existing embedding benchmarks primarily evaluate textual interactions on video-text and audio-text pairs, and visual interactions on image-text pairs, they fail to capture the full spectrum of omni-modal inputs and to support both unimodal and multimodal interactions. To evaluate audio interactions as well as video-audio alignment, we introduce OmniCHOIR, an omni-interactive text-video-audio-to-audio (TVA2A) retrieval benchmark that systematically evaluates textual, visual, and audio interactions. In OmniCHOIR, given a video with mixed audio and a query (which may consist of a single modality or a combination of modalities), the task is to retrieve the sound that

[2]We treat text as an interaction prompt (p^t) unless otherwise specified, due to its ambiguous role.

matches the intended target from the mixture, conditioned on the corresponding background context. To further increase compositional and reasoning difficulty, we construct multiple hard negative distractors with different background sounds, the same background but containing sounds from the same target category, and both different sound categories and backgrounds. This design establishes a unique property of OmniCHOIR: it evaluates embedding models not only across diverse interaction modalities, but also under fine-grained compositional and reasoning challenges, providing a more rigorous and realistic testbed for omni-interactive embeddings.

In summary, our contributions are three-fold:

- **Method Novelty:** We propose OmniUE, the first omni-interactive omni-modal universal embedder that accepts text, visual, and audio prompts as interactions jointly with arbitrary combinations of text, visual, and audio inputs via segmenters and multi-layer context aggregation to generate a unified embedding. The visual segmenter allows users to provide different types of visual prompts, and the audio segmenter enables users to provide temporal audio spans.
- **Benchmark Novelty:** Since there is no omni-interactive embedding benchmark, we introduce OmniCHOIR, composed of omni-interactive TVA2A retrieval that aims to assess embedders' reasoning, compositional, and unimodal or multimodal user-conditioned capabilities. OmniCHOIR enables advancements on omni-modal interaction scenarios that remain unexplored.
- **Experiment Novelty:** OmniUE significantly and consistently outperforms state-of-the-art embedding baselines across multiple benchmarks, including 18 text-interactive video tasks (MMEB-v2-video), 30 text-interactive audio tasks (MAEB), 5 visual-interactive text-image-to-text tasks (SCaR), and the omni-interactive OmniCHOIR benchmark, achieving average improvements of 10.4%, 1.0%, 83.1%, and 21.9%, respectively.

2 Related Works

LLM-Based Multimodal Embedders. Most existing works focus on advancing MLLM-based embedders [26, 27, 69, 35, 34], which demonstrate effective performance and instruction-following capabilities for handling diverse user queries in embedding tasks, outperforming traditional dual-encoder paradigms, e.g., [47, 60, 31, 67, 11, 68, 55, 21]. In parallel, audio-language representation learning has been advanced by works such as [16], which introduce CLAP to align audio and text in a shared embedding space for classification and retrieval. Subsequent efforts further improve performance by scaling training data and enhancing caption quality for fine-grained audio-text alignment [57, 37, 3, 66]. While these methods advance the realm of multimodal representation learning, the primary hurdles lie in their focus on text paired with a single additional modality and in their reliance on text as the sole interface for user-conditioned interaction. Recently, omni-modal embedders have been proposed to unify visual (image and video), audio, and text within a shared embedding space, either through binding modalities [20, 73] or omni-modal LLMs [58, 64, 9, 36, 50, 25]; nonetheless, these approaches continue to rely on text as the sole interaction modality. To date, [53] is the only work that extends interaction beyond text by introducing visual prompts; yet, it is limited to the image-text setting and does not generalize to video or audio, nor does it support audio-based interactions. To address these limitations, our proposed OmniUE is able to accommodate not only omni-modal inputs, but also any individual modality (or their combinations) as interaction media by introducing additional video and audio segmenters and context aggregation that aims to incorporate complex information.

Interactive Embedding Benchmarks. With the rapid advancement of MLLM-based embedders for diverse embedding tasks, evaluation benchmarks have shifted from traditional unimodal (e.g., BEIR [51], MTEB [39]) and bimodal retrieval settings to large-scale, instruction-driven multimodal benchmarks that assess instruction-following capabilities across diverse scenarios. Meng et al. [38] introduce MMEB-v2-video, comprising 18 text-video embedding tasks spanning video QA, classification, retrieval, and moment retrieval, by reformulating both retrieval and generative tasks into embedding tasks with different textual instructions. MAEB [2] provides a large-scale text-audio embedding benchmark covering 30 tasks across speech, music, environmental sounds, and cross-modal reasoning in over 100 languages. As existing benchmarks primarily rely on textual instructions, SCaR [53] introduces a large-scale visual-interactive text-image-to-text (TI2T) retrieval benchmark, composed of five visual grounding datasets for region-level caption retrieval within global scene context. While we leverage MMEB-v2-video, MAEB, and SCaR to extensively evaluate textual and visual interactive embedding capabilities, these benchmarks do not assess scenarios involving

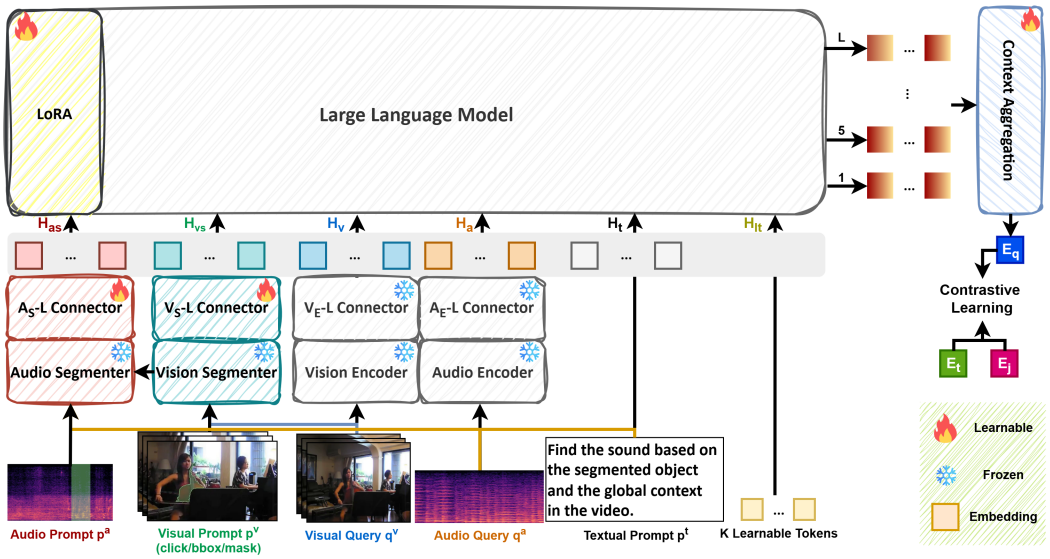

Figure 2: OmniUE overview. Compared to existing omni-LLM-based embedders (shown in gray; e.g., [58]), OmniUE introduces three key components: 1) audio and visual segmenters that enable emerging interaction capabilities; 2) additional learnable tokens for modeling complex and compositional multimodal inputs; and 3) a context aggregation module that selectively weighs the token-layer importance of contextual tokens to produce the final embedding for contrastive learning.

video-audio alignment or audio-based interactions, where audio spans themselves serve as prompts. To fill this gap, we introduce OmniCHOIR, an omni-interactive text-video-audio-to-audio (TVA2A) retrieval benchmark that challenges models with unimodal and multimodal interactions.

3 OmniUE

3.1 Overview

Fig. 2 presents an overview of our proposed OmniUE, which supports text-only, video-only (q^v), audio-only (q^a), and joint multimodal inputs, along with visual-, audio-, and textual interaction prompts (p^v, p^a, p^t). Visual and audio inputs, including both queries and interaction prompts, are processed by their respective segmenters/encoders and connectors to produce user-conditioned holistic hidden representations. In addition, K learnable tokens are appended to the input sequence and jointly processed by the LLM to capture compositional and complex multimodal information, which has been found to be useful for the image-text paradigm [34]. The hidden states of these K tokens are extracted at a fixed interval of l layers and aggregated via a context aggregation module to produce the final embedding E. The query (E_q), target (E_t), and negative (E_j) embeddings are then jointly optimized using the InfoNCE loss [52]. We adopt GradCache [18] to enable larger batch sizes and improve generalization over in-batch negatives, following prior work [53].

Converting modality inputs into embeddings. The omni-LLM consists of vision and audio encoders and their corresponding connectors. Given a multimodal input (either a query or a target) that comprises any combination of text, visual, and audio modalities, the resulting tokenized d-dimensional embeddings for each modality are denoted as H_v, H_a, and H_t, respectively.

Appending additional multi-token inputs. A common practice to obtain a single representation E is to use the final hidden state of the EOS token from the last layer, as is widely adopted in MLLM embedders [38, 53, 58]. However, naively relying on the last-layer representation of a single token may degrade the inherent capabilities of MLLMs [41], and a single token is often insufficient to encode complex omni-modal information. To extend the multi-token design to the compositionally denser omni-modal realm, we augment each input sequence with $K = 4$ learnable tokens $H_{lt} \in \mathbb{R}^{K \times d}$. These tokens are randomly initialized and trained to capture embedding-specific

information beyond the original MLLM vocabulary. Nevertheless, this design remains restricted to using hidden states from the final layer; we address this limitation in Sec. 3.3.

3.2 Enabling Emerging Visual- and Audio-Interactive Capabilities

While H_v and H_a encode holistic representations of visual and audio inputs, they are insufficient for incorporating user-specified interaction prompts and capturing fine-grained information (e.g., entity-level cues for vision and temporal span-level cues for audio). To address this limitation, we introduce two additional segmenter streams for vision and audio, which support points, bounding boxes, masks, text, and temporal spans as interaction inputs, producing the corresponding interaction embeddings H_{vs} and H_{as}.

Encoding interactive prompts into interaction embeddings. We adopt pretrained SAM-3 [6] as the vision segmenter since it effectively segments entities based on user-specified regions of interest. In parallel, SAM-Audio [49] serves as the audio segmenter, enabling sound separation from mixed audio conditioned on optional video, text, or audio temporal spans. A key advantage is that SAM-3 can be viewed as a submodule of SAM-Audio, allowing them to be integrated into a unified interaction framework. Specifically, we extract the visual feature map $F_{vs} \in \mathbb{R}^{|F_{vs}^H| \times |F_{vs}^W| \times d_{vs}}$ from before the mask prediction head in SAM-3, which integrates video inputs from the image encoder and visual prompts from the detector. Similarly, the audio feature map $F_{as} \in \mathbb{R}^{|F_{as}| \times d_{as}}$ is obtained before the decoder (i.e., after the Diffusion Transformer [43] and the iterative sampling procedure) in SAM-Audio. We use feature maps prior to decoding since they already encode interaction-aware information from both the inputs and interaction prompts, eliminating the need for additional decoding back into embedding space. For non-visual interactive scenarios, visual prompts are omitted, directly yielding the corresponding feature maps. For non-audio interactive scenarios, prompts are set to their default forms, namely zero vectors for videos, `<null>` tokens for spans, and empty strings for text, following the design of SAM-Audio. Therefore, segmenter-derived embeddings are employed even in purely text-interactive settings, enabling the model to capture entity-level cues while preserving the global context from the vision encoder within the omni-LLM.

Connecting interaction embeddings to the unified LLM space. Directly flattening visual and audio feature maps leads to significantly long sequences when aligning with the LLM (e.g., $|F_{vs}^H| \times |F_{vs}^W| = 288 \times 288$ in SAM-3). To this end, we introduce segmenter connectors by applying a Conv2D layer to reduce the spatial resolution of visual feature maps and a Conv1D layer to project audio feature maps into the LLM-compatible sequence space. These representations are then passed through h-head self-attention connectors (SA) and MLP to produce the final interaction embeddings:

$$H_{vs} = \text{MLP}_{vs}(\text{SA}_{vs}(\text{Conv2D}(F_{vs}))) \in \mathbb{R}^{|F'_{vs}| \times d}, H_{as} = \text{MLP}_{as}(\text{SA}_{as}(\text{Conv1D}(F_{as}))) \in \mathbb{R}^{|F'_{as}| \times d}, \quad (1)$$

where $|F'_{vs}|$ and $|F'_{as}|$ are the compressed lengths of H_{vs} and H_{as}, respectively.

3.3 Aggregating Contextual and Intermediate Tokens to the Final Embedding

A key challenge in omni-modal embedding is how to effectively aggregate heterogeneous signals, especially for multimodal inputs and user interactions. Relying solely on the final-layer representation, as in prior work (e.g., [58, 41]), overlooks multi-level or modality-specific information captured at intermediate depths. In contrast, MLP-based layer aggregation methods (e.g., [50]) introduce substantial parameter overhead due to the high dimensionality of LLM hidden states, with a complexity of $O(L \cdot d^2)$, where L is the number of layers. For instance, fine-tuning Qwen2.5-omni-3B (d is 2048 and L is 36) [62] with LoRA requires around 45M trainable parameters, whereas the MLP used for late fusion alone introduces 155M additional parameters, potentially leading to underfitting (see Appendix E.1).

To address this, we first construct a unified sequence by concatenating holistic context, user-conditioned representations, and learnable tokens as $H = [H_{as}, H_{vs}, H_v, H_a, H_t, H_{lt}]$, which is then fed into the LLM. To capture multi-level representations, we extract the hidden states of the learnable tokens H_{lt} every l layers (with $l = 4$) starting from the first (i.e., layers 1, 5, 9, ...), and include the final layer if not already selected. Then, we introduce a context aggregation module to fuse these representations. Let $Z \in \mathbb{R}^{M \times K \times d}$ denote the collected K token features across the M selected layers. Instead of treating all layers equally, we introduce a set of normalized layer-token importance weights $W \in \mathbb{R}^{M \times K}$, allowing the model to learn a global preference over representation

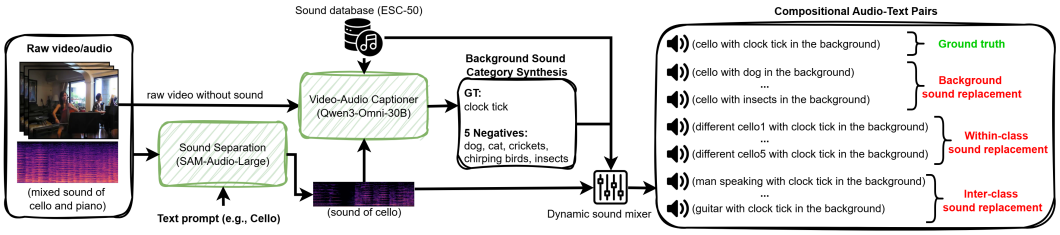

Figure 3: The data collection pipeline for OmniCHOIR. We use SAM-Audio-Large for sound separation conditioned on text prompts, and Qwen3-Omni-30B as a captioner to synthesize ground-truth and negative background sound categories based on the visual information from the silent video. A dynamic sound mixer then injects background sounds into the target audio track with randomized temporal intervals and volume levels (-15 dB to -5 dB).

levels and tokens, inspired by [44]. The final embedding $E \in \mathbb{R}^d$ is:

$$E = \phi(\tilde{Z}), \tilde{Z} = \sum_{m=1}^{M} \tilde{W}_m \odot Z_m \in \mathbb{R}^{K \times d}, \tilde{W} = \text{softmax}(W), \tilde{W}_m \in R^K, \tag{2}$$

where $\odot$ denotes element-wise multiplication with broadcasting along the feature dimension, ϕ is mean pooling over the K tokens, and the softmax operation is applied across both the layer and token dimensions. In this manner, the model selectively integrates contextual and optionally interaction-aware signals from both local and global representations retrieved from earlier and later LLM layers, while maintaining computational efficiency.

4 OmniCHOIR: Omni-Interactive TVA2A Retrieval Benchmark

4.1 Overview

As discussed in Sec. 2, existing benchmarks primarily focus on textual-interactive embedding evaluation, while only [53] introduces a visual-interactive benchmark, which is limited to the text-image paradigm. To complement existing benchmarks and move toward a more comprehensive evaluation, we introduce OmniCHOIR, an omni-interactive text-video-audio-to-audio retrieval benchmark. We focus on TVA2A retrieval because audio mixtures inherently involve the superposition of multiple sources, leading to ambiguities that are qualitatively more challenging than visual occlusion, and thus require precise multimodal interaction (e.g., temporal spans or cross-modal cues) to resolve. Moreover, compared to visual region-based interaction, audio-centric interaction remains unexplored in existing benchmarks. OmniCHOIR challenges models not only in reasoning and compositionality, but also in handling diverse combinations of multimodal interaction prompts. Each instance in OmniCHOIR comprises a raw video, its synchronized mixed audio, visual segmentation masks for the target subject, audio temporal spans corresponding to the target sound, and textual descriptions of the desired sound.

Problem Definition. Given a raw video V, the corresponding mixed audio A, and an interaction prompt P, where P can be any combination of p^a, p^v, and p^t, the goal is to retrieve the most relevant audio that matches both the target sound and its background context from a set of 16 audio candidates C: $c_y = \text{argmax}_{c_i \in C} \cos(\theta(V, A, P), \theta(c_i)), C = \{c_1, \ldots, c_{16}\}$, where θ denotes the embedding model. In practice, textual instructions are always included (i.e., *"Find the sound from the mixed audio and the video: {text condition}"* for p^t, *"Sound occurring between {start}s and {end}s"* for p^a, and *"Identify the sound from the highlighted region"* for p^v), and are omitted here for simplicity.

4.2 Collection Pipeline

Fig. 3 illustrates the pipeline for constructing OmniCHOIR. We build upon SAM-Audio-Bench[3] [49] as the base dataset, which contains 819 10-second samples with 24 fps video V, audio A, and multi-modal interaction prompts P. SAM-Audio-Bench is a real-world multimodal separation benchmark

[3]It is not included when training SAM-Audio.

Table 1: Results on 18 MMEB tasks, with scores averaged per meta-task. Improvements are computed against the best-performing 2B/3B and 7B baselines. We highlight the **best** and second-best in each column. All results are from the leaderboard and [9], except for LCO-Emb, which we reproduced.

	CLS	QA	RET	MRET	**Overall**
Model ↓ | #Datasets →	5	5	5	3	18
Text + Image + Video					
VLM2Vec-v2 (Qwen2-VL-2B) [38]	39.3	34.3	28.8	38.5	34.9
UME-R1 (Qwen2-VL-2B) [29]	44.3	51.2	32.9	39.7	42.2
UniME-V2 (LLaVA-OneVision-7B) [22]	37.2	50.6	28.9	39.6	39.0
UME-R1 (Qwen2-VL-7B) [29]	48.6	60.7	38.2	39.3	47.5
CAFe (LLaVa-OneVision-7B) [65]	35.8	58.7	34.4	39.5	42.4
Text + Image + Video + Audio (Omni-modal)					
Omni-Embed-Nemotron (Qwen2.5-omni-3B) [64]	40.5	44.3	32.7	25.6	36.9
LCO-Emb (Qwen2.5-omni-3B) [58]	42.9	56.7	30.0	43.8	43.3
e5-omni (Qwen2.5-omni-3B) [9]	40.2	48.5	33.2	40.7	40.6
LCO-Emb (Qwen2.5-omni-7B) [58]	39.3	57.6	24.8	26.5	38.2
e5-omni (Qwen2.5-omni-7B) [9]	46.6	52.9	36.7	34.2	43.5
WAVE (Qwen2.5-omni-7B) [50]	50.7	45.9	34.7	39.5	43.1
Ours (Omni-modal)					
OmniUE-3B disabling audio & vision segmenters	47.8	52.6	34.5	45.1	45.0
OmniUE-7B disabling audio & vision segmenters	54.7	59.9	37.6	35.5	48.2
OmniUE-3B	51.1	57.8	36.9	**47.7**	48.4
OmniUE-7B	**57.8**	**62.9**	**40.9**	41.3	**51.8**
Improvements (3B)	+6.8	+1.1	+3.7	+3.9	+5.1
Improvements (7B)	+9.2	+2.2	+2.7	+1.7	+4.3

with diverse taxonomic coverage, including speech cleaning, speaker separation, music removal, instrument stems (37 classes), and general environmental sounds. Each instance provides human-annotated visual masklets (for on-screen sounding objects), temporal spans, and textual descriptions as interaction signals P. The video and audio data are sourced from six public datasets: AudioSet [19], VGGSound [7], MUSIC [70], MUSIC-AVQA [30], AVSpeech [17], and CondensedMovies [4]. Importantly, all samples in SAM-Audio-Bench are held out, ensuring that OmniUE was not exposed to these data during training and thereby supporting evaluation of zero-shot generalization.

Since SAM-Audio-Bench does not provide clean target sounds, we apply SAM-Audio-Large [49] to perform sound separation, followed by manual validation. After filtering out invalid samples and missing videos, we retain 479 instances. Next, we employ Qwen3-Omni-30B [63] as a captioner, conditioned on the silent video and the target sound, to generate one ground-truth and five negative background-sound categories based on the ESC-50 sound taxonomy [46]. The prompt instructs the model to select plausible background sounds for the scene that are not strongly supported by the visual context and are semantically distinct from a ground-truth category. We also encourage Qwen3-Omni to produce diverse and creative negatives, ensuring coverage across different categories, while reducing the risk of overfitting to narrow patterns.

We then introduce a dynamic sound mixer, which combines the target sound with the ground-truth background sound to form the reference audio. The background sounds are randomly sampled from the ESC-50 dataset based on the synthesized categories and injected at randomized temporal intervals. To ensure acoustic diversity and realism, we vary the relative volume levels of the background sounds within a range of -15 dB to -5 dB relative to the target signal. To construct challenging negative candidates, we design three replacement strategies, each producing five candidates to balance difficulty and diversity: **1) Background sound replacement:** replace only the background sound using the five generated negatives. **2) Within-class sound replacement:** replace the target sound with another separated sound from the same category within SAM-Audio-Bench. **3) Inter-class sound replacement:** replace the target sound with another target sound from a different category. All corresponding text descriptions, which serve as OmniCHOIR's text conditions, are generated using the template *"{targeted sound} with {background sound} in the background."*, forming the final instances of OmniCHOIR. Additional prompt templates and examples are provided in Appendix D.

Table 2: Results on 30 MAEB tasks. Scores are averaged per meta-task. Improvements are computed relative to the best-performing 3B and 7B baselines. We highlight the **best** and second-best results in each column. Baseline results are from the MAEB leaderboard.

	Clf	M.Clf	PC	Rrnk	Clust	A. Rtrvl	X. Rtrvl	Zero Clf.	**Overall**
Model ↓ | #Datasets →	10	2	3	1	3	1	8	2	30
Two-Tower-Based Embedders									
microsoft/msclap-2023 [16]	45.0	5.8	53.6	75.4	**15.2**	87.3	9.4	12.6	31.1
laion/larger_clap_general [57]	51.7	2.3	51.9	66.8	6.6	93.2	9.8	14.9	32.2
LLM-Based Embedders									
LCO-Emb (Qwen2.5-Omni-3B) [58]	56.4	41.6	66.7	75.4	1.3	67.7	**50.3**	62.2	50.7
Qwen2-Audio-7B [12]	**62.7**	10.7	56.9	80.8	12.7	33.9	1.6	12.4	33.7
LCO-Emb (Qwen2.5-Omni-7B) [58]	58.0	45.7	67.3	78.7	1.7	78.2	**50.3**	64.5	52.2
Ours (Omni-modal)									
OmniUE-3B disabling audio & vision segmenters	53.6	20.5	67.3	87.0	1.3	80.4	44.7	56.5	45.1
OmniUE-7B disabling audio & vision segmenters	59.0	23.5	68.5	88.3	1.8	92.4	24.5	66.6	45.5
OmniUE-3B	57.3	35.3	67.5	**88.4**	1.7	81.1	46.5	63.5	50.5
OmniUE-7B	59.1	**49.9**	**68.8**	88.2	2.7	**93.3**	47.7	**67.1**	**53.4**
Improvements (3B)	+0.9	-	+0.8	+13.0	-	+13.4	-	+1.3	-
Improvements (7B)	-	+4.2	+1.5	+7.4	-	+15.1	-	+2.6	+1.2

5 Experiments

5.1 Experimental Setup

Benchmarks. To comprehensively evaluate OmniUE's capabilities, we conduct experiments on textual, visual, and omni-interactive benchmarks. Specifically, **MMEB-v2-video** [38] assesses 18 textual-interactive video tasks, including classification (CLR), question answering (QA), retrieval (RET), and moment retrieval (MRET). **MAEB** [2] examines 30 textual-interactive audio tasks, including classification (Clf), multi-label classification (M.Clf), pairwise classification (PC), reranking (Rrnk), clustering (Clust), audio retrieval (A.Rtrvl), cross-modal retrieval (X.Rtrvl), and zero-shot classification (Zero Clf.). **SCaR** [53] evaluates 5 visual-interactive text-image-to-text (TI2T) retrieval tasks across different domains. In addition, our proposed **OmniCHOIR** assesses omni-interactive TVA2A retrieval tasks that consist of text, visual regions, and audio temporal spans as interaction media. We use the same metrics as the official benchmarks, and Recall@1 is used for OmniCHOIR.

Implementation Details. We use 3B and 7B checkpoints of Qwen2.5-Omni [62] as the omni-LLM backbone for OmniUE-3B and OmniUE-7B, respectively, both of which utilize sam-audio-base [49] equipped with SAM-3 [6] as the audio and vision segmenters. LoRA [24] is applied for the LLM with the rank set to 32 and alpha set to 64, and the audio prompt and visual prompt connectors are trained from scratch to align the segmenter space to the omni-LLM space. OmniUE is trained on around 3.5M multimodal samples from public datasets, with a batch size of 1024. To evaluate whether the interaction streams remain beneficial even when no explicit visual or audio interaction prompts are available, disabling the visual & audio segmenters denotes the corresponding interaction streams being disabled only during inference (i.e., no inputs to visual and audio interaction streams), while the model is still trained with them enabled with default inputs as described in Sec. 3.2. To evaluate the emerging visual- and audio-interactive capabilities, we do not include the corresponding interactions in the training datasets, i.e., only textual-interactive datasets are used for training OmniUE. Detailed datasets and configurations are provided in Appendix C, with ablation and parameter studies, and inference time and memory analyses in Appendices E.1 and E.2, respectively.

5.2 Textual-Interactive Performance

Video benchmarks (MMEB-v2-video). Tab. 1 summarizes the overall performance on MMEB-v2-video. OmniUE consistently outperforms all baselines, including text+image+video and omni-modal models, across all four meta-task scenarios. Quantitatively, OmniUE-3B surpasses existing 2B/3B models by 5.1 points on average, while OmniUE-7B improves over the 7B baselines by 4.3 points. To analyze the source of these gains, we compare OmniUE with variants that remove the audio and visual segmenter streams. The performance drop in these variants indicates that incorporating fine-grained, entity-level information is critical for textual-interactive embedding. Notably, the segmenter streams remain beneficial even when no explicit interaction prompts are provided, suggesting that they complement global multimodal representations with localized details via segmenter connectors.

Table 3: Results on 5 SCaR tasks. Improvements are computed relative to the best-performing 2B and 7B baselines. We highlight the **best** and second-best results in each column. Baseline results are from the original SCaR paper.

Model ↓ | **Datasets** →	RefCOCO+	RefCOCOg	VisualGenome	COCO-Stuff	ADE20K	**Overall**
Textual-Interactive-Only Models						
VLM2Vec-2B [27]	24.5	29.5	22.3	19.4	24.6	24.1
VLM2Vec-7B [27]	23.2	29.1	14.7	25.0	22.4	22.9
MMRet-7B [72]	27.1	21.8	15.2	26.0	22.6	22.5
UniME-7B [22]	31.4	32.8	19.0	25.3	23.0	26.3
Textual + Visual-Interactive Models						
VIRTUE-2B [53]	28.8	42.4	24.4	29.9	27.5	30.6
VIRTUE-7B [53]	33.0	35.3	19.6	27.1	23.8	27.8
Omni-Interactive Models (Ours)						
OmniUE-3B	64.4	57.5	**36.8**	57.0	41.0	51.3
OmniUE-7B	**65.3**	**66.2**	35.5	**61.6**	**47.5**	**55.2**
Improvements (3B)	+35.6	+15.1	+12.4	+27.1	+13.5	+20.7
Improvements (7B)	+32.3	+30.9	+15.9	+34.5	+23.7	+27.4

Table 4: Results on OmniCHOIR. We highlight the **best** results in each column. Text refers to the interaction text prompt. For span scenarios, we 1) convert the selected span into text using the same instructions as in OmniUE and 2) directly crop the corresponding audio segment, reported as text/crop.

Model ↓ | **Conditions** →	Text	Span	Mask	Text+Span	Text+Mask	Span+Mask	Text+Span+Mask
Two-Tower-Based Models							
laion/larger_clap_general [57]	16.1	1.7/12.7	-	12.7	-	-	-
Textual-Interactive Omni-Modal Models							
ImageBind [20]	16.8	3.3/13.5	-	18.1	-	-	-
LCO-Emb-3B [58]	19.6	18.7	-	20.5	-	-	-
LCO-Emb-7B [58]	22.3	22.4	-	23.8/23.8	-	-	-
WAVE-7B [50]	20.0	23.1/23.4	-	24.5/24.7	-	-	-
Omni-Interactive Omni-Modal Models (Ours)							
OmniUE-3B	23.6	25.3	25.6	23.6	25.7	25.7	26.4
OmniUE-7B	**25.5**	**26.3**	**26.7**	**27.4**	**26.7**	**27.6**	**28.4**

Audio benchmarks (MAEB). Beyond video evaluation, Tab. 2 reports results on 30 textual-interactive audio tasks from MAEB. OmniUE consistently outperforms both two-tower and LLM-based embedders, achieving substantial gains across all meta-tasks. On average, OmniUE-3B achieves on par performance with LCO-Emb-3B. Compared to LCO-Emb-7B, OmniUE achieves the best performance in 7 of the 8 MAEB meta-task categories, with particularly large gains of +15.1 on Audio Retrieval, +9.5 on Re-ranking, and +4.2 on Multi-label Classification. Ablation results further show that removing the segmenter streams leads to consistent degradation in performance. This confirms that segmenter-enhanced representations provide critical fine-grained information for audio-text interaction. Additional experiments on textual-interactive video-to-audio, audio-to-video, video-audio-to-text, and text-video-audio are included in Appendix E.4.

5.3 Visual-Interactive TI2T Performance (SCaR)

To evaluate visual-interactive capabilities, we compare OmniUE with state-of-the-art visual-interactive models on the SCaR benchmark, where the goal is to find the most relevant text caption based on a given image, text instruction, and a bounding box. As shown in Tab. 3, OmniUE consistently outperforms all baselines across all five datasets. Specifically, OmniUE achieves significant improvements of 20.7 and 27.4 points over the strongest 2B and 7B baselines, respectively. These results demonstrate that OmniUE generalizes beyond textual-interactive scenarios to visual-interactive settings. In addition to training on omni-modal datasets, we attribute these gains to its unified design, which integrates global context with interaction prompts and leverages context aggregation with learnable tokens to capture fine-grained information.

5.4 Omni-Interactive TVA2A Performance (OmniCHOIR)

To further evaluate omni-interactive capabilities, we stress-test OmniUE on the proposed OmniCHOIR benchmark, which includes seven uni and multimodal interaction conditions, i.e., text, span, mask, text+span, text+mask, span+mask, text+span+mask. As OmniCHOIR requires omni-modal inputs,

we compare against two-tower-based and textual-interactive omni-modal models that only accept text-based interaction prompts. As shown in Tab. 4, OmniUE already outperforms all baselines under the text-only condition, which we attribute to the audio segmenter's ability to leverage textual cues to focus on user-specified audio. In particular, both visual (masks) and audio (temporal span) interactions individually outperform the text-only baselines. The deleterious performance of converting spans with text or cropped intervals demonstrates that the gains of OmniUE cannot be explained by simple span-cropping or text-conversion heuristics, but arise from the proposed omni-interaction representation. Moreover, combining multiple interaction modalities leads to additional improvements, and using all available modalities yields the best performance. This trend highlights a key property of OmniUE: it not only supports heterogeneous interaction modalities, but also effectively composes them to refine retrieval. These results demonstrate that OmniUE can flexibly leverage both individual and combined user interactions, enabling unified omni-interactive retrieval beyond text-only paradigms.

6 Conclusion

In this paper, we propose OmniUE, the first omni-interactive universal embedder that jointly encodes omni-modal inputs and omni-modal interactions as holistic context and fine-grained guidance. Distinct from existing approaches that support only text and image as interaction modalities, OmniUE leverages off-the-shelf SAM-3 and SAM-Audio as the visual and audio segmenters with multi-token layer aggregation, enabling the model to produce user-conditioned query embeddings and unlock emerging capabilities in visual and audio interactive embedding scenarios. To comprehensively evaluate these capabilities, we introduce OmniCHOIR, an omni-interactive TVA2A retrieval benchmark that incorporates text, visual masks, and audio spans as interaction media, along with challenging negative candidates constructed by altering background sounds or target sound categories. Extensive experiments across textual, visual, audio, and joint interactive embedding benchmarks demonstrate that OmniUE not only outperforms existing methods on their respective tasks, but also enables new functionalities under diverse combinations of interaction modalities. We believe OmniUE serves as a generic framework for omni-interactive, omni-modal embeddings, while OmniCHOIR offers a new evaluation paradigm for assessing such capabilities, together paving the way toward universal embedders. Further discussions on limitations and broader impacts are in Appendix A.

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *CoRR*, abs/1609.08675, 2016.

[2] A. E. Assadi, I. Chung, C. Xiao, R. Solomatin, A. Jha, R. Chand, S. Singh, K. Wang, A. S. Khan, M. M. Nasser, S. Fong, P. He, A. Xiao, A. S. Munot, A. Shrivastava, A. Gazizov, N. Muennighoff, and K. C. Enevoldsen. MAEB: massive audio embedding benchmark. *CoRR*, abs/2602.16008, 2026.

[3] J. Bai, H. Liu, M. Wang, D. Shi, W. Wang, M. D. Plumbley, W.-S. Gan, and J. Chen. Audiosetcaps: Enriched audio captioning dataset generation using large audio language models. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024. URL `https://openreview.net/forum?id=uez4PMZwzP`.

[4] M. Bain, A. Nagrani, A. Brown, and A. Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *ACCV (5)*, Lecture Notes in Computer Science, pages 460–479. Springer, 2020.

[5] D. Bolya, P. Huang, P. Sun, J. H. Cho, A. Madotto, C. Wei, T. Ma, J. Zhi, J. Rajasegaran, H. Rasheed, J. Wang, M. Monteiro, H. Xu, S. Dong, N. Ravi, D. Li, P. Dollár, and C. Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *CoRR*, abs/2504.13181, 2025.

[6] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Rädle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. Hazra, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H. K. Cheng, P. Dollár, N. Ravi, K. Saenko,

P. Zhang, and C. Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

[7] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, pages 721–725. IEEE, 2020.

[8] H. Chen, L. Wang, N. Yang, Y. Zhu, Z. Zhao, F. Wei, and Z. Dou. mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. In *ACL (Findings)*, volume ACL 2025 of *Findings of ACL*, pages 8254–8275. Association for Computational Linguistics, 2025.

[9] H. Chen, S. Gao, R. Timofte, T. Sakai, and Z. Dou. e5-omni: Explicit cross-modal alignment for omni-modal embeddings. *CoRR*, abs/2601.03666, 2026.

[10] T. Chen, A. Siarohin, W. Menapace, E. Deyneka, H. Chao, B. E. Jeon, Y. Fang, H. Lee, J. Ren, M. Yang, and S. Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, pages 13320–13331. IEEE, 2024.

[11] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829. IEEE, 2023.

[12] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou. Qwen2-audio technical report. *CoRR*, abs/2407.10759, 2024.

[13] X. Cong, H. Yang, A. Wang, Y. Wang, Y. Yang, C. Zhang, and C. Ma. Viva: Vlm-guided instruction-based video editing with reward optimization. *arXiv preprint arXiv:2512.16906*, 2025.

[14] K. Drossos, S. Lipping, and T. Virtanen. Clotho: an audio captioning dataset. In *ICASSP*, pages 736–740. IEEE, 2020.

[15] Y. Du, W. Huang, D. Zheng, Z. Wang, S. Montella, M. Lapata, K.-F. Wong, and J. Z. Pan. Rethinking memory in llm based agents: Representations, operations, and emerging topics, 2025. URL https://arxiv.org/abs/2505.00675.

[16] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang. CLAP learning audio concepts from natural language supervision. In *ICASSP*, pages 1–5. IEEE, 2023.

[17] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37(4):112, 2018.

[18] L. Gao, Y. Zhang, J. Han, and J. Callan. Scaling deep contrastive learning batch size under memory limited setup. In *RepL4NLP@ACL-IJCNLP*, pages 316–321. Association for Computational Linguistics, 2021.

[19] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, pages 776–780. IEEE, 2017.

[20] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. Imagebind one embedding space to bind them all. In *CVPR*, pages 15180–15190. IEEE, 2023.

[21] T. Gu, K. Yang, Z. Feng, X. Wang, Y. Zhang, D. Long, Y. Chen, W. Cai, and J. Deng. Breaking the modality barrier: Universal embedding learning with multimodal llms. *CoRR*, abs/2504.17432, 2025.

[22] T. Gu, K. Yang, K. Zhang, X. An, Z. Feng, Y. Zhang, T. W. Cai, J. Deng, and L. Bing. Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning. *CoRR*, abs/2510.13515, 2025.

[23] M. Han, L. Yang, X. Chang, L. Yao, and H. Wang. Shot2story: A new benchmark for comprehensive understanding of multi-shot videos. In *ICLR*. OpenReview.net, 2025.

[24] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022.

[25] C. Huynh, M. Luong, and A. Shrivastava. Efficient and high-fidelity omni modality retrieval. *CoRR*, abs/2603.02098, 2026.

[26] T. Jiang, M. Song, Z. Zhang, H. Huang, W. Deng, F. Sun, Q. Zhang, D. Wang, and F. Zhuang. E5-V: universal embeddings with multimodal large language models. *CoRR*, abs/2407.12580, 2024.

[27] Z. Jiang, R. Meng, X. Yang, S. Yavuz, Y. Zhou, and W. Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *ICLR*. OpenReview.net, 2025.

[28] C. D. Kim, B. Kim, H. Lee, and G. Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT (1)*, pages 119–132. Association for Computational Linguistics, 2019.

[29] Z. Lan, L. Niu, F. Meng, J. Zhou, and J. Su. UME-R1: exploring reasoning-driven generative multimodal embeddings. *CoRR*, abs/2511.00405, 2025.

[30] G. Li, Y. Wei, Y. Tian, C. Xu, J. Wen, and D. Hu. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, pages 19086–19096. IEEE, 2022.

[31] J. Li, D. Li, C. Xiong, and S. C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022.

[32] K. Liang, J. Kruk, S. Qian, X. Yang, S. Bi, Y. Yao, S. Nie, M. Zhang, L. Liu, J. F. Fisac, S. Zhou, and S. Hosseini. Learning personalized agents from human feedback, 2026. URL `https://arxiv.org/abs/2602.16173`.

[33] J. Liu, S. Chen, X. He, L. Guo, X. Zhu, W. Wang, and J. Tang. VALOR: vision-audio-language omni-perception pretraining model and dataset. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(2): 708–724, 2025.

[34] Q. Liu, X. Liang, Z. Zhang, Z. Qing, F. Zhou, Y. Chen, X. Tang, Y. Hu, and P. Henderson. Rematch: Boosting representation through matching for multimodal retrieval. *CoRR*, abs/2511.19278, 2025.

[35] Y. Liu, Y. Zhang, J. Cai, X. Jiang, Y. Hu, J. Yao, Y. Wang, and W. Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *CVPR*, pages 4015–4025. Computer Vision Foundation / IEEE, 2025.

[36] X. Ma, L. Gao, S. Zhuang, J. S. Zhan, J. Callan, and J. Lin. Tevatron 2.0: Unified document retrieval toolkit across scale, language, and modality. In *SIGIR*, pages 4061–4065. ACM, 2025.

[37] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:3339–3354, 2024.

[38] R. Meng, Z. Jiang, Y. Liu, M. Su, X. Yang, Y. Fu, C. Qin, Z. Chen, R. Xu, C. Xiong, Y. Zhou, W. Chen, and S. Yavuz. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *CoRR*, abs/2507.04590, 2025.

[39] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. MTEB: massive text embedding benchmark. In *EACL*, pages 2006–2029. Association for Computational Linguistics, 2023.

[40] OpenClaw. Openclaw — personal ai assistant. `https://github.com/openclaw/openclaw`, 2025. Accessed: March 2026.

[41] Y. Ouali, A. Bulat, A. Xenos, A. Zaganidis, I. M. Metaxas, B. Martínez, and G. Tzimiropoulos. Vladva: Discriminative fine-tuning of lvlms. In *CVPR*, pages 4101–4111. Computer Vision Foundation / IEEE, 2025.

[42] S. Ouyang, J. Yan, I. Hsu, Y. Chen, K. Jiang, Z. Wang, R. Han, L. T. Le, S. Daruki, X. Tang, V. Tirumalashetty, G. Lee, M. Rofouei, H. Lin, J. Han, C. Lee, and T. Pfister. Reasoningbank: Scaling agent self-evolving with reasoning memory. *CoRR*, abs/2509.25140, 2025.

[43] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4172–4182. IEEE, 2023.

[44] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *NAACL-HLT*, pages 2227–2237. Association for Computational Linguistics, 2018.

[45] Philipp Schmid. The new skill in ai is not prompting, it's context engineering. `https://www.philschmid.de/context-engineering`, 2025. Accessed: March 2026.

[46] K. J. Piczak. ESC: dataset for environmental sound classification. In *ACM Multimedia*, pages 1015–1018. ACM, 2015.

[47] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[48] N. Ravi, V. Gabeur, Y. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C. Wu, R. B. Girshick, P. Dollár, and C. Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR*. OpenReview.net, 2025.

[49] B. Shi, A. Tjandra, J. Hoffman, H. Wang, Y.-C. Wu, L. Gao, J. Richter, M. Le, A. Vyas, S. Chen, C. Feichtenhofer, P. Dollár, W.-N. Hsu, and A. Lee. Sam audio: Segment anything in audio, 2025. URL `https://arxiv.org/abs/2512.18099`.

[50] C. Tang, Q. Xiao, K. Mei, T. Wang, F. Rao, and C. Zhang. WAVE: Learning unified & versatile audio-visual embeddings with multimodal LLM. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=MiV3WXDYJb`.

[51] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS Datasets and Benchmarks*, 2021.

[52] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[53] W.-Y. Wang, K. Tateishi, Q. Wu, S. Takahashi, and Y. Mitsufuji. VIRTUE: Visual-interactive text-image universal embedder. In *ICLR*, 2026. URL `https://openreview.net/forum?id=H4RgGzx4iL`.

[54] X. Wang, J. Wu, J. Chen, L. Li, Y. Wang, and W. Y. Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, pages 4580–4590. IEEE, 2019.

[55] C. Wei, Y. Chen, H. Chen, H. Hu, G. Zhang, J. Fu, A. Ritter, and W. Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *ECCV (87)*, volume 15145 of *Lecture Notes in Computer Science*, pages 387–404. Springer, 2024.

[56] Q. Wu, S. Cui, S. Hayakawa, W. Wang, H. Wakaki, and Y. Mitsufuji. MCA: modality composition awareness for robust composed multimodal retrieval. *CoRR*, abs/2510.15543, 2025.

[57] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, pages 1–5. IEEE, 2023.

[58] C. Xiao, H. P. Chan, H. Zhang, W. Xu, M. Aljunied, and Y. Rong. Scaling language-centric omnimodal representation learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=eA3Aum0rpA`.

[59] B. Xu. AI agent systems: Architectures, applications, and evaluation. *CoRR*, abs/2601.01743, 2026.

[60] H. Xu, G. Ghosh, P. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP (1)*, pages 6787–6800. Association for Computational Linguistics, 2021.

[61] J. Xu, T. Mei, T. Yao, and Y. Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296. IEEE Computer Society, 2016.

[62] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, B. Zhang, X. Wang, Y. Chu, and J. Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

[63] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu, Y. Lv, Y. Wang, D. Guo, H. Wang, L. Ma, P. Zhang, X. Zhang, H. Hao, Z. Guo, B. Yang, B. Zhang, Z. Ma, X. Wei, S. Bai, K. Chen, X. Liu, P. Wang, M. Yang, D. Liu, X. Ren, B. Zheng, R. Men, F. Zhou, B. Yu, J. Yang, L. Yu, J. Zhou, and J. Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

[64] M. Xu, W. Zhou, Y. Babakhin, G. de Souza Pereira Moreira, R. Ak, R. Osmulski, B. Liu, E. Oldridge, and B. Schifferer. Omni-embed-nemotron: A unified multimodal retrieval model for text, image, audio, and video. *CoRR*, abs/2510.03458, 2025.

[65] H. Yu, Z. Zhao, S. Yan, L. Korycki, J. Wang, B. He, J. Liu, L. Zhang, X. Fan, and H. Yu. CAFE: unifying representation and generation with contrastive-autoregressive finetuning. In *ICCVW*, pages 6345–6356. IEEE, 2025.

[66] Y. Yuan, D. Jia, X. Zhuang, Y. Chen, Z. Chen, Y. Wang, Y. Wang, X. Liu, X. Kang, M. D. Plumbley, and W. Wang. Sound-vecaps: Improving audio generation with visually enhanced captions. In *ICASSP*, pages 1–5. IEEE, 2025.

[67] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11941–11952. IEEE, 2023.

[68] K. Zhang, Y. Luan, H. Hu, K. Lee, S. Qiao, W. Chen, Y. Su, and M. Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. In *ICML*. OpenReview.net, 2024.

[69] X. Zhang, Y. Zhang, W. Xie, M. Li, Z. Dai, D. Long, P. Xie, M. Zhang, W. Li, and M. Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *CVPR*, pages 9274–9285. Computer Vision Foundation / IEEE, 2025.

[70] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. H. McDermott, and A. Torralba. The sound of pixels. In *ECCV (1)*, Lecture Notes in Computer Science, pages 587–604. Springer, 2018.

[71] B. Zheng, M. Y. Fatemi, X. Jin, Z. Z. Wang, A. Gandhi, Y. Song, Y. Gu, J. Srinivasa, G. Liu, G. Neubig, and Y. Su. Skillweaver: Web agents can self-improve by discovering and honing skills. *CoRR*, abs/2504.07079, 2025.

[72] J. Zhou, Y. Xiong, Z. Liu, Z. Liu, S. Xiao, Y. Wang, B. Zhao, C. J. Zhang, and D. Lian. Megapairs: Massive data synthesis for universal multimodal retrieval. In *ACL (1)*, pages 19076–19095. Association for Computational Linguistics, 2025.

[73] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, H. Wang, Y. Pang, W. Jiang, J. Zhang, Z. Li, C. Zhang, Z. Li, W. Liu, and L. Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*. OpenReview.net, 2024.

A Limitation and Broader Impacts

Limitations. While OmniUE represents a significant step toward reinforcing omni-interactive capabilities in embedding models, the main limitation lies in its dependence on pre-trained models (e.g., Qwen2.5-omni and SAM-Audio) since it is infeasible to train all modules from scratch. Xiao et al. [58] introduce a generation-representation scaling law showing that embedding performance is positively related to generative performance, suggesting that our proposed method can be boosted by incorporating future state-of-the-art pretrained models. Nonetheless, this important and orthogonal direction reflects our overarching objective: to pave the way for embedding models that can process any kind of user intent, while maintaining robust universal embedding performance. In addition, there is a tradeoff between computational efficiency and interaction precision when using masks as interaction media (see Appendix E.2); however, we provide further analysis using bounding boxes as alternative efficient visual-interactive solutions. In future work, we aim to explore more use scenarios for OmniUE (e.g., interactive similarity metrics), advance our framework, and build other complementary combinations as interactive benchmarks, to fully unlock the potential of omni-interactive universal embedders. Moreover, we plan to scale OmniCHOIR to a larger collection.

Broader Impacts. As an omni-interactive, omni-modal embedder, OmniUE can facilitate the development of interaction-aware models (e.g., [13]) by providing effective user-conditioned embeddings. In addition, OmniUE can serve as a similarity-based evaluation model, similar to prior approaches such as ImageBind [20]. However, unlike conventional methods that rely on holistic similarity, OmniUE enables interaction-aware similarity by jointly modeling multimodal inputs and user-provided interaction prompts. Finally, OmniUE can be integrated into agentic pipelines that rely on contextual embeddings for applications such as memory, personalization, and adaptive decision-making, where its interaction-aware representations enable more fine-grained and controllable behavior. As a representation learning framework, OmniUE poses a lower direct risk of generating harmful or deceptive content compared to generative multimodal models. However, its fine-grained interaction-aware embeddings could potentially be misused in downstream applications. For instance, the integration of multimodal user signals necessitates careful attention to data privacy to prevent the unintended encoding of sensitive personal attributes.

B LLM Usage

In addition to Qwen3-Omni used for building the OmniCHOIR benchmark (i.e., sound category selection), we use LLMs solely for polishing the manuscript.

C Implementation Details

C.1 Parameter Setting

OmniUE-3B and OmniUE-7B are trained with a τ of 0.02 and a learning rate of 2e-5. The OmniUE 3B and 7B training was conducted on 8$\times$H100 80GB, which took around 177 and 195 hours, respectively. The remaining configurations are identical to the default settings of Qwen2.5-omni, SAM-Audio, and SAM-3. The iterative sampling procedure is the ODE solver following the original SAM-Audio paper. While SAM-Audio originally uses 16 steps, we found it performs on par with 8 steps, while reducing GPU memory usage. The kernel size and stride of Conv2D and Conv1D are set to 4. Detailed configurations are summarized in Tab. 5.

C.2 Training Datasets

Tab. 6 summarizes the data sources used and their corresponding modalities, resulting in a total of approximately 3.5M multimodal samples. To enrich modality coverage, we use Qwen/Qwen3-Omni-30B-A3B-Instruct [63] to synthesize text captions from paired video and audio inputs, where the generated captions jointly capture both visual and auditory information. For VGGSound, we adopt text captions from AudioSetCaps [3]. Our early experiments suggest that incorporating these synthesized captions improves embedding performance on MMEB-v2-video and MAEB. It is important to note that the training datasets used in OmniUE do not overlap with the OmniCHOIR samples.

Table 5: Detailed configurations for OmniUE 3B and 7B models.

Configurations	OmniUE-3B	OmniUE-7B
Omni-LLM	Qwen2.5-Omni-3B	Qwen2.5-Omni-7B
Vision Segmenter	facebook/sam3	
Audio Segmenter	facebook/sam-audio-base	
# of heads in SAC (h)	2	
$\lvert F_{vs}^{H} \rvert \times \lvert F_{vs}^{W} \rvert$	288×288	
$\lvert F_{as} \rvert$	maximum 250	
$\lvert F_{vs}' \rvert$	maximum 256	
$\lvert F_{as}' \rvert$	maximum 128	
d_{vs}	256	
d_{as}	128	
Max sequence length	20480	
Max frame resolution	156800	
Iterative sampling procedure	8-step ODE solver	
A_S-L connector & V_S-L connector	Random initialized	
Audio sampling rate	16kHz	
LoRA rank	32	
LoRA dropout	0.1	
LoRA alpha	64	
Temperature τ	0.02	
Batch size	1024	
Training steps	3300	
Warmup steps	200	
Learning rate	2e-5	
d	2048	3584
l	4	
K	4	
GPU	8×H100 80G	
Precision	bf16	
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)	
Training time	177 hours	195 hours

Table 6: Overview of the training data used to train OmniUE. * denotes synthesized text captions.

Data Source	Modalities	# of Samples
AudioSet [19]	Video, Audio, Text*	86k
WavCaps [37]	Audio, Text	404k
MSR-VTT [61]	Video, Text	180k
PE-Videos [5]	Video, Text	983k
VALOR [33]	Video, Audio, Text	2k
VATEX [54]	Video, Text	5k
AudioCaps [28]	Audio, Text	45k
Clotho [14]	Audio, Text	19k
Panda70m [10]	Video, Audio, Text*	278k
Shot2Story [23]	Video, Audio, Text	970k
VGGSound [7]	Video, Audio, Text*	183k
Youtube8m [1]	Video, Audio, Text*	269k

D OmniCHOIR Details

D.1 OmniCHOIR Examples

For better readability, we present OmniCHOIR examples in the supplementary material (i.e., index.html), where it can be seen that the benchmark cover diverse domains.

D.2 Prompt Template for Video-Audio Captioning

We summarize the prompt template for Qwen3-Omni-30B below.

Prompt Template for Qwen3-Omni

You are an expert audio-visual scene analyst. Given a video and its primary audio, infer one background sound that naturally fits the visible scene, and generate several distractor background sounds that are plausible in the same broad environment but are not supported by the video and are semantically different from the gold sound. Follow the scene strictly: do not invent unlikely entities or actions that are not visible or strongly implied.
The primary sound in this video is: "original_caption".
Below is the complete list of allowed environmental sound categories. You MUST choose all outputs ONLY from this list, using exact string matches:esc50_categories
Task:
1. Select exactly ONE category from the allowed list that would plausibly occur as a background sound in this video's scene.
2. Select exactly FIVE other categories from the same allowed list that are plausible in the broad environment, but are not strongly supported by the video context and are semantically different from the gold category.
3. Be creative and diverse negative categories.
Strict constraints:
- Every output must be an exact category name copied from the allowed list.
- Do NOT output any sound label outside the list, even if it seems more natural.
- Do NOT output the primary sound itself.
- Do NOT output synonyms, paraphrases, or near-duplicates of a listed category.
- DO NOT use music_playing that is not in the list, use the closest plausible allowed background category instead.
- Negative backgrounds must be realistic distractors in the same broad scene, not absurdly impossible sounds.
- Negative backgrounds must be different from the gold category and from each other.
Respond ONLY with valid JSON (no extra text):

```
{
    "gold_background": "<exact category from list>",
    "negative_backgrounds":
        [
            "<exact category from list>",
            "<exact category from list>",
            "<exact category from list>",
            "<exact category from list>",
            "<exact category from list>"
        ]
}
```

E Additional Experiments

E.1 Ablation and Parameter Study

To analyze the relative contributions of different components and parameters, we conduct ablation and parameter studies on 15 variant designs and choices. To iteratively refine the results, all studies here are experimented with a batch size of 512 and 3000 training steps on OmniUE-3B, which take around 48 hours for each training configuration. The evaluation on OmniCHOIR uses text as conditions.

Table 7: Ablation study on the impact of each proposed component. "w/o visual and audio streams" denotes the removal of visual and audio interactions during both training and testing stages. "w/o visual stream" and "w/o audio stream" denote the removal of visual-only and audio-only interactions during both stages. "w/o multi-token inputs" replaces the K learnable tokens with the EOS hidden state, reducing W to $\mathbb{R}^M$. "w/o layer aggregation" removes the context aggregation module and applies mean pooling over the K tokens from the last layer.

	MMEB-v2-video	MAEB	SCaR	OmniCHOIR
OmniUE-3B	44.7	47.8	50.5	22.2
w/o visual stream	42.7	47.6	32.0	20.5
w/o audio stream	44.1	47.4	49.8	18.7
w/o visual and audio streams	42.5	47.5	30.9	16.3
w/o multi-token inputs	43.8	46.6	47.2	20.1
w/o layer aggregation	43.2	45.0	49.8	18.5

Relative contributions of each module. As shown in Tab. 7, removing any proposed component consistently degrades performance across all interaction benchmarks, demonstrating that each module provides complementary benefits. In particular, removing the visual or audio interaction streams leads to substantial drops on SCaR and OmniCHOIR, confirming that these streams are critical for modality-specific interactions. Furthermore, removing the visual stream negatively impacts MMEB-v2-video and SCaR, while removing the audio stream hurts more on MAEB; both removal settings are inferior to OmniCHOIR compared to OmniUE. Overall, the results establish that 1)

multi-token inputs and 2) layer aggregation yield consistent gains across all benchmarks, while 3) interaction streams provide targeted improvements in their corresponding visual and audio scenarios.

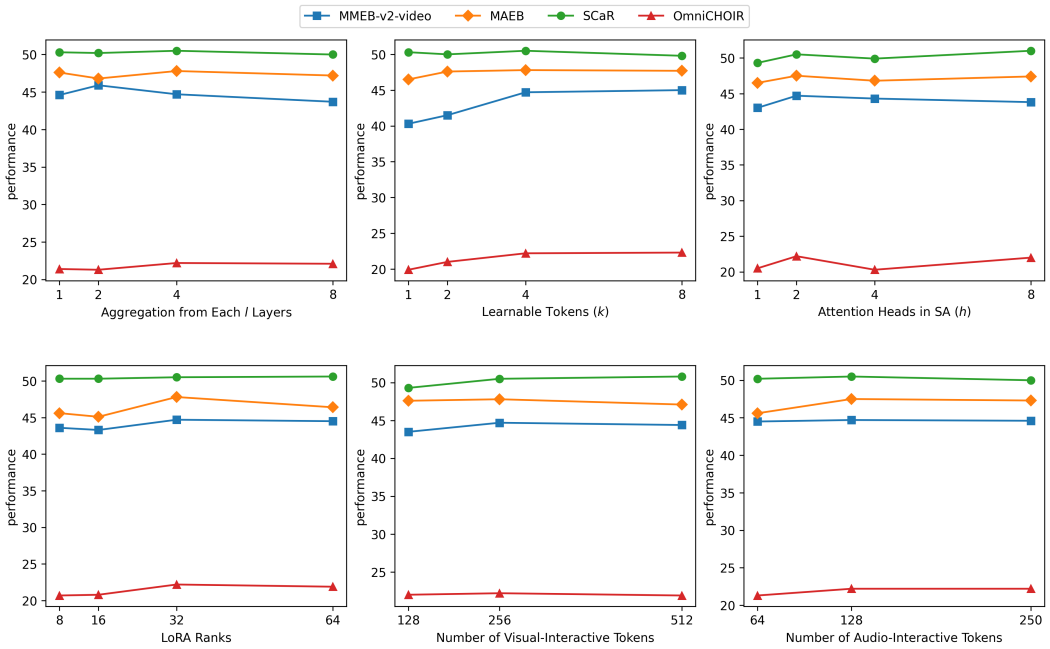

Figure 4: Parameter study on 1) Aggregation from each l layers; 2) # of learnable tokens (k); 3) # of attention heads in SA (h); 4) LoRA ranks; 5) number of visual-interactive tokens ($|F'_{vs}|$); and 6) number of audio-interactive tokens ($|F'_{as}|$).

Relative contributions of omni-interactions and architectural changes. To further isolate the contribution of our proposed components, we additionally perform two complementary full-training ablations: 1) removing the proposed multi-token inputs and layer aggregation (architectural refinements), and 2) removing the proposed visual and audio interaction streams (omni-interaction formulation). As shown in Tab. 8, removing the interaction streams causes the largest degradation on interaction-centric benchmarks (SCaR: -18.4, OmniCHOIR: -7.1), demonstrating that the proposed omni-interaction formulation is the primary factor enabling new interaction capabilities. Meanwhile, removing multi-token inputs and layer aggregation consistently reduces performance across benchmarks, confirming that these architectural designs further improve embedding quality. Together, these results suggest that OmniUE's improvements arise from the proposed omni-interaction formulation and embedding design built on top of pretrained models, rather than solely from the pretrained modules.

Table 8: Ablation study on the omni-interactions and architectural changes.

	MMEB-v2-video	MAEB	SCaR	OmniCHOIR
OmniUE-3B	48.4	50.5	51.3	26.4
w/o learnable tokens & context aggregation	43.4	44.7	46.3	23.5
w/o visual & audio streams	46.5	48.2	32.9	19.3

Impact of key hyperparameters. To evaluate the robustness of OmniUE, we perform a parameter study over six factors: 1) the aggregation interval l across LLM layers, 2) the number of learnable tokens K, 3) the number of attention heads in the self-attention (SA) module, 4) different LoRA ranks, 5) the number of visual-interactive tokens ($|F'_{vs}|$), and 6) the number of audio-interactive tokens ($|F'_{as}|$). As shown in Fig. 4, aggregating more intermediate hidden states generally improves performance compared to sparse aggregation (e.g., $l = 8$), with the best overall performance achieved at $l = 4$, indicating a balance between information richness and redundancy. For learnable tokens, using a single token ($K = 1$) significantly degrades performance, whereas increasing to $K = 2$ consistently improves results, suggesting that multiple tokens are necessary to capture heterogeneous

multimodal interactions. In addition, using 2 and 8 attention heads yields comparable performance; we adopt 2 heads as the default configuration due to its more favorable performance–efficiency trade-off. Finally, the ablations on LoRA rank, visual-interactive tokens, and audio-interactive tokens further validate our design choices. Varying the number of visual-interactive tokens primarily affects performance on visual benchmarks, while varying the number of audio-interactive tokens predominantly impacts audio benchmarks. This modality-aligned sensitivity confirms that the proposed token design effectively captures modality-specific interactions and that our selected configurations achieve a balanced and robust performance across tasks.

Table 9: Ablation study with OmniUE-3B on 1) audio segmenter size, 2) visual segmenter version, 3) segmenter connector design, and 4) layer aggregation variants. The highlighted row indicates the configurations used in OmniUE-3B and OmniUE-7B.

Choice	MMEB-v2-video	MAEB	SCaR	OmniCHOIR
small	43.4	46.7	49.3	21.5
base	44.7	47.8	50.5	22.2
large	44.2	48.1	50.2	20.8

Choice	MMEB-v2-video	MAEB	SCaR	OmniCHOIR
SAM-2.1-B+	43.9	44.5	46.1	20.7
SAM-3	44.7	47.8	50.5	22.2

Alternative	MMEB-v2-video	MAEB	SCaR	OmniCHOIR
w/o SA	42.8	45.1	47.9	20.5
w/ SA	44.7	47.8	50.5	22.2

Alternative	MMEB-v2-video	MAEB	SCaR	OmniCHOIR
MLP	41.8	47.0	45.9	20.2
Layer-token weights	44.7	47.8	50.5	22.2

Design choices for variant modules. To further analyze key design decisions, Tab. 9 compares several alternatives: 1) the pretrained audio segmenter, 2) the pretrained visual segmenter, 3) the role of self-attention (SA) in segmenter-related connectors, and 4) different aggregation strategies. For segmenters, the base variant of SAM-Audio consistently achieves the overall best performance among available sizes, and SAM-3 outperforms SAM2.1 [48] as visual segmenters. This trend indicates that stronger pretrained segmenters directly translate to improved embedding quality in multimodal interaction tasks. Removing self-attention (SA) from the connectors leads to consistent performance degradation, demonstrating its importance. This suggests that SA enables the model to selectively attend to different segmentation tokens conditioned on user intent, thereby improving interaction modeling. Finally, comparing aggregation methods shows that a standard MLP underperforms our proposed layer-token importance weighting. We attribute this to the substantially larger number of trainable parameters in MLP-based aggregation, which introduces optimization bias and reduces generalization, whereas our design provides a more structured and efficient aggregation mechanism.

E.2 Analysis on Latency and Memory

Table 10: Analysis on inference time and memory for unimodal and multimodal query modes of OmniUE-7B on OmniCHOIR.

	Query Mode	Time (per sample)	Memory	Performance
		OmniCHOIR		
LCO-Emb-7B	Text	1.0s	17.7GB	22.3
WAVE-7B	Text	1.3s	18.2GB	20.0
OmniUE-7B (Ours)	Text	1.3s	20.8GB	25.5
	Span	1.3s	20.8GB	26.3
	Mask	9.4s	20.8GB	26.7
	Text+Span	1.3s	20.8GB	27.4
	Text+Mask	9.3s	20.8GB	26.7
	Span+Mask	9.4s	20.8GB	27.6
	Text+Span+Mask	10.5s	20.8GB	28.4
	Bbox	1.5s	20.8GB	26.4
	Text+Bbox	1.5s	20.8GB	26.7
	Text+Span+Bbox	2.1s	20.8GB	27.0

We report the average end-to-end inference latency (s) and GPU memory consumption (GB) across all seven operational modes of OmniUE-7B and LCO-Emb-7B on the OmniCHOIR benchmark. As illustrated in Tab. 10, while OmniUE introduces a marginal overhead to facilitate omni-interactive capabilities, it achieves a substantial 14.3% performance gain under text-conditioned settings. To

further optimize the inference pipeline, we investigate the trade-off between segmentation masks and bounding boxes (Bbox). While substituting masks with Bboxes results in a slight performance penalty, it significantly reduces latency, which provides a practical recipe for balancing computational throughput with high-fidelity performance in real-world applications.

E.3 Robustness of OmniUE to Interaction Prompts

To evaluate the robustness of OmniUE to interactions, we conducted additional experiments to analyze the effects of imperfect interaction inputs, including 1) randomly sampled visual masks, 2) randomly sampled audio spans, and 3) removing the text prompt completely.

Table 11: Analysis on noisy masks, spans, and text interaction prompts.

Model ↓ | **Conditions →**	Text	Span	Mask	Text+Span	Text+Mask	Span+Mask	Text+Span+Mask
OmniUE-7B	**25.5**	**26.3**	**26.7**	**27.4**	**26.7**	**27.6**	**28.4**
1) Random masks	-	-	16.8	-	21.3	22.5	22.9
2) Random spans	-	19.6	-	19.8	-	20.5	23.7
3) w/o text prompts	10.7	-	-	16.9	21.0	-	22.3

E.4 Additional Modality Directions with Textual Interactions

To further support the universal embedder claim, we additionally evaluate video-to-audio (V2A) and audio-to-video (A2V) retrieval on VGGSound, and video-audio-to-text (VA2T) and text-to-video-audio on VALOR-32K, using the standard Recall@1 metric. We compare against the strongest method applicable to each benchmark (WAVE-7B [50] on VGGSound and VALOR [33] on VALOR-32K). As shown in Tab. 12, the results show that OmniUE-7B outperforms the state-of-the-art methods on both benchmarks, further strengthening the capability toward any-to-any retrieval.

Table 12: Performance on V2A/A2V on VGGSound and VA2T/T2VA on VALOR-32K.

	V2A	A2V
WAVE-7B	25.0	25.8
OmniUE-7B	27.2	26.1

	VA2T	T2VA
VALOR	74.5	73.2
OmniUE-7B	78.9	79.3