

# OmniRetriever: Any-to-Any Audio-Video-Text Retrieval via Fusion-as-Teacher Distillation

Yunze Liu   Chi-Hao Wu   Enmin Zhou   Junxiao Shen\*

Memories.ai Research

Project Page: [OmniRetriever](#)

## Abstract

Unified multimodal embedding spaces have become the standard interface for cross-modal retrieval and multimodal RAG, and recent audio-video-text (AVT) encoders extend this setting to three modalities. Such encoders can produce a joint  $(T, V, A)$  embedding whenever all three modalities are available, but standard pairwise InfoNCE objectives leave this signal unused during training. We close this gap with **fusion-as-teacher distillation**, which treats a stop-gradient copy of the fused embedding as a teacher signal for the single-modal embedding, paired with a Tuple-InfoNCE term that supervises the fused embedding directly. We instantiate this objective as **OmniRetriever-7B**. Across six zero-shot retrieval benchmarks, *OmniRetriever-7B* reaches R@1 of 19.14 on Clotho text-to-audio and 47.60 on MSR-VTT text-to-video. To stress-test joint representations, we further release **OmniRetriever-Bench**, a 12-direction AVT retrieval benchmark totaling 3,782 triples; on it *OmniRetriever-7B* attains AVG-all 34.84, improving over the best prior open-source AVT method by +8.03. Model weights, datasets, and code are released.

## 1 Introduction

Cross-modal retrieval relies on encoders that map queryable modalities into a single shared embedding space. The canonical CLIP recipe (Radford et al., 2021) trains two separate towers and aligns them with pairwise InfoNCE. Recent omni-modal systems extend this to three modalities of  $\{T, V, A\}$  and split into two architectural families. *Multi-encoder* systems such as ImageBind (Girdhar et al., 2023) and LanguageBind (Zhu et al., 2024a) keep separate per-modality encoders and align them pairwise to a fixed image or language anchor. *Unified* encoders such as WAVE (Tang et al., 2025) and Omni-Embed-Nemotron (Xu

et al., 2025) instead route all three modalities through one backbone and produce a joint embedding  $\mathbf{z}_{TVA}$  on a single joint forward pass.

Unified encoders are particularly attractive for AVT retrieval because  $\mathbf{z}_{TVA}$  is exactly the representation that a dual-modal or full-tuple query (e.g.,  $T+V \rightarrow A$  or  $A+T+V \rightarrow T$ ) needs at inference. Yet the standard training recipe for unified AVT encoders never invokes this joint forward as a supervision signal: WAVE and Omni-Embed-Nemotron both optimise three pairwise InfoNCE losses on the single-modal sub-encoders only, and the joint output is computed only at inference time.

The result is that the single-modal sub-encoders of current unified AVT systems receive no direct signal from the representation formed when all modalities are present. Empirically, the largest performance gaps appear on audio-anchored retrieval. On Clotho  $A \rightarrow T$ , the best open omni-modal system, Omni-Embed-Nemotron, reaches  $R@1 = 3.5$ , while CLAP-family audio-text specialists (Wu et al., 2023; Mei et al., 2024; Niizumi et al., 2025) trained on  $(T, A)$  pairs alone reach 25 to 26 on the same direction. This pattern motivates joint-representation supervision; it does not by itself identify a data-quality or routing cause. We therefore evaluate all twelve any-to-any directions over  $\{T, V, A\}$  (six single-modal, six dual-modal) that a practical AVT retriever has to serve.

We close this gap by using the joint embedding itself as the supervision signal. **Fusion-as-teacher distillation**  $\mathcal{L}_D$  takes a stop-gradient copy of  $\mathbf{z}_{TVA}$  as the teacher for the single-modal sub-encoders: each of  $\mathbf{z}_T$ ,  $\mathbf{z}_V$ ,  $\mathbf{z}_A$  is pulled toward  $\mathbf{z}_{TVA}$  by InfoNCE. Because the teacher is the same backbone consumed jointly rather than an external encoder, the sub-encoders receive the cross-modal context that no unimodal teacher provides, at the cost of one additional joint forward pass per step. A complementary Tuple-InfoNCE refinement (Liu et al., 2021, 2020)  $\mathcal{L}_T$  supervises

\*Corresponding author.

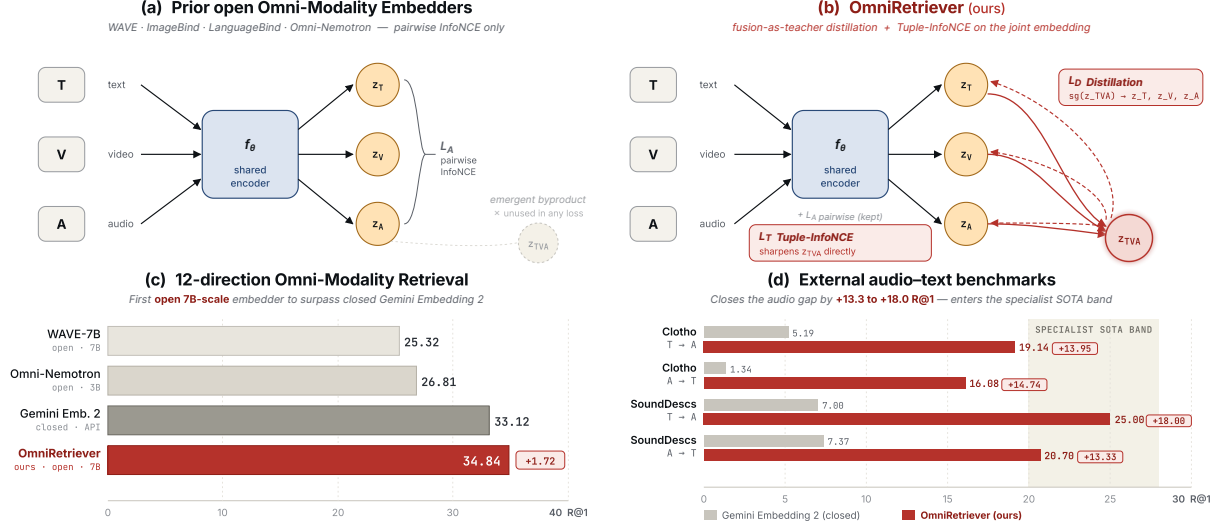


Figure 1: **Method overview.** *OmniRetriever* uses the joint embedding  $\mathbf{z}_{TVA}$ , which is unused by pairwise training (a), as a supervision target (b) via fusion-as-teacher distillation  $\mathcal{L}_D$  and a Tuple-InfoNCE term  $\mathcal{L}_T$ . This yields a new open result on 12-direction AVT retrieval (c) and a 13.3 to 18.0 R@1 gain over Gemini Embedding 2 on external audio-text benchmarks (d).

$\mathbf{z}_{TVA}$  itself with modality-cycled hard negatives, preventing the joint vector from collapsing onto the strongest pair gradient (in practice  $T-V$ ).

We instantiate this recipe as *OmniRetriever-7B*, an open 7 B AVT retriever. On six standard zero-shot retrieval benchmarks, *OmniRetriever-7B* reaches R@1 of 19.14 on Clotho  $T \rightarrow A$ , 25.00 on SoundDescs  $T \rightarrow A$ , and 47.60 on MSR-VTT  $T \rightarrow V$ . To probe the six dual-modal directions ( $T \leftrightarrow AV$ ,  $A \leftrightarrow TV$ ,  $V \leftrightarrow AT$ ) that no public retrieval benchmark currently evaluates, we additionally release *OmniRetriever-Bench*, a 12-direction AVT retrieval pool of 3,782 held-out triples. On *OmniRetriever-Bench*, *OmniRetriever-7B* reaches AVG-all 34.84, +8.03 over Omni-Embed-Nemotron. A cross-backbone replication (Table 8) reproduces the dominant  $\mathcal{L}_D$  contribution at smaller scale, indicating that the recipe is not tied to a particular backbone.

### Contributions.

- Fusion-as-teacher distillation: the joint multi-modal embedding  $\mathbf{z}_{TVA}$  of a unified AVT encoder is used to supervise its own single-modal sub-encoders.  $\mathcal{L}_D$  alone gives the dominant single-loss gain; the Tuple-InfoNCE term  $\mathcal{L}_T$  further improves the  $A \leftrightarrow V$  routes.
- *OmniRetriever-7B*, an open 7 B AVT retriever evaluated zero-shot on six standard audio-text and video-text benchmarks, with controlled external loss ablations on Clotho and MSR-VTT.
- *OmniRetriever-Bench*, a 12-direction AVT re-

trieval pool of 3,782 held-out triples covering all six single- and six dual-modal directions, the first public benchmark to evaluate dual-modal AVT queries.

## 2 Related Work

### Pairwise contrastive vision-language alignment.

CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) establish image-text InfoNCE as the standard recipe. Subsequent work scales the data pipeline (Schuhmann et al., 2022; Fang et al., 2024; Xu et al., 2024), replaces softmax with the sigmoid loss (Zhai et al., 2023; Tschannen et al., 2025), and re-anchors with vision foundation models (Siméoni et al., 2025; Assran et al., 2025; Fini et al., 2025). A parallel line of work turns MLLMs into encoders (Wang et al., 2024a; BehnamGhader et al., 2024; Lee et al., 2025). Multimodal extensions (VLM2Vec (Jiang et al., 2025), GME (Zhang et al., 2025), MM-Embed (Lin et al., 2025), Qwen3-VL-Embedding (Li et al., 2026)) fine-tune MLLMs with contrastive supervision over LLM-synthesized triplets. All apply pairwise InfoNCE per modality combination. Our PAIRWISE baseline follows this recipe.

**Omni-modal embedding.** ImageBind (Girdhar et al., 2023) and LanguageBind (Zhu et al., 2024a) align  $N \geq 3$  modalities pairwise to a fixed image or language anchor. AudioCLIP (Guzhov et al., 2022) and VATT (Akbari et al., 2021) optimize

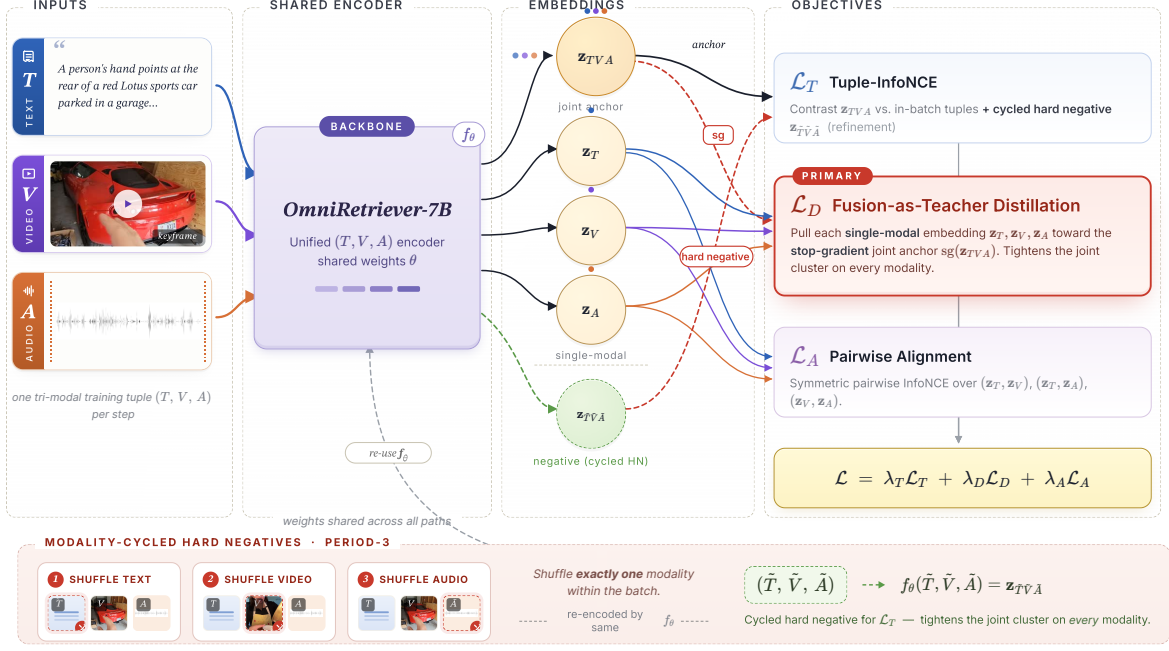


Figure 2: **OmniRetriever training overview.** A shared encoder  $f_\theta$  consumes the three modalities jointly, producing the full-modal anchor  $\mathbf{z}_{TVA}$ , or individually, producing  $\mathbf{z}_T, \mathbf{z}_V, \mathbf{z}_A$ .  $\mathcal{L}_D$  (fusion-as-teacher distillation, primary; Section 3.2) pulls each single-modality embedding toward a stop-gradient copy of  $\mathbf{z}_{TVA}$ .  $\mathcal{L}_T$  (Tuple-InfoNCE refinement; Section 3.3) contrasts  $\mathbf{z}_{TVA}$  with an independently augmented joint view and a modality-cycled corrupted-tuple pool  $\mathbf{z}_{\tilde{T}\tilde{V}\tilde{A}}$  (Equation (4)).  $\mathcal{L}_A$  (pairwise alignment; Section 3.1) ties pairs of single-modality embeddings via symmetric InfoNCE. At each step the hard negative perturbs one of  $T, V, A$  on a period-3 schedule.

three pairwise InfoNCE losses over video, audio, and text. VAST (Chen et al., 2023b) adds a 4-way alignment objective on video–audio–subtitle–caption quadruples. WAVE (Tang et al., 2025) extends Qwen2.5-Omni with hierarchical visual fusion and a dual audio encoder but still adds two-modality InfoNCE losses, as does Omni-Embed-Nemotron (Xu et al., 2025). In all of these systems, the joint multimodal embedding is produced as a side effect of pairwise training and never used as a supervision target inside the same backbone.

**Masked reconstruction and teacher construction.** Multimodal masked autoencoders such as MultiMAE and M3AE learn cross-modal representations by reconstructing masked inputs or features (Bachmann et al., 2022; Geng et al., 2022). PHG-MAE instead masks entire modalities and can aggregate predictions from multiple random masking paths (Pîrvu and Leordeanu, 2025). Related hypergraph systems combine predictions along multiple task pathways into pseudo-label teachers (Marcu et al., 2023; Pîrvu et al., 2023). These methods and ours share the use of multimodal context for self-supervision. *OmniRe-*

*triever* neither reconstructs masked content nor ensembles outputs from repeated masking runs: it performs retrieval-space self-distillation, using the same encoder’s fused tuple embedding as a stop-gradient teacher for its single-modal representations and directly contrasting augmented fused views.

### 3 Method

Let  $\mathcal{D} = \{x_i\}$  be an AVT dataset where each  $x_i = (x_i^{(T)}, x_i^{(V)}, x_i^{(A)})$  carries text, video, and audio. A unified embedder  $f_\theta : \mathcal{X} \rightarrow \mathbb{R}^d$  produces single-modal embeddings  $\mathbf{z}_i^{(m)} = f_\theta(x_i^{(m)})$  and a joint embedding  $\mathbf{z}_i^{(TVA)} = f_\theta(T_i, V_i, A_i)$  from the same forward backbone. Figure 2 overviews the three training objectives.

#### 3.1 Preliminary: Pairwise Alignment ( $\mathcal{L}_A$ )

Existing open AVT embedders (Girdhar et al., 2023; Zhu et al., 2024a; Tang et al., 2025; Xu et al., 2025) optimize a symmetric InfoNCE loss

over modality pairs:

$$\mathcal{L}_{\text{NCE}}^{(m,m')} = -\frac{1}{B} \sum_i \log \frac{e^{\mathbf{z}_i^{(m)\top} \mathbf{z}_i^{(m')}/\tau}}{\sum_j e^{\mathbf{z}_i^{(m)\top} \mathbf{z}_j^{(m')}/\tau}}, \quad (1)$$

$$\mathcal{L}_A = \sum_{(m,m') \in \mathcal{P}} \mathcal{L}_{\text{NCE}}^{(m,m')}, \quad \mathcal{P} = \{(T, V), (T, A), (V, A)\}.$$

The PAIRWISE baseline we report throughout uses  $\mathcal{L}_A$  alone. Since  $\mathcal{L}_A$  never operates on the joint  $(T, V, A)$  vector,  $\mathbf{z}_{TVA}$  is neither supervised nor distilled: pairwise alignment binds the three streams only to the extent that each pair enforces in isolation. *OmniRetriever* retains  $\mathcal{L}_A$  and adds two joint-level losses,  $\mathcal{L}_D$  and  $\mathcal{L}_T$ , described next.

### 3.2 Fusion-as-Teacher Distillation ( $\mathcal{L}_D$ )

The joint forward of  $f_\theta$  on  $(T, V, A)$  is the only step at which all three modalities interact in the network. We use  $\mathbf{z}_{TVA}$  as the teacher. A symmetric InfoNCE pulls each single-modal embedding toward a stop-gradient copy of  $\mathbf{z}_{TVA}$  and pushes it away from the joint vectors of other in-batch samples:

$$\mathcal{L}_D = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathcal{L}_{\text{NCE}}^{(m, \text{sg}(\mathbf{z}_{TVA}))}, \quad (2)$$

with  $\mathcal{M} = \{T, V, A\}$ . The teacher  $\text{sg}(\mathbf{z}_{TVA})$  already encodes the three modalities jointly, so each single-modal student is trained against the joint geometry that an  $A+T+V$  query reaches at inference. Teacher and student share  $f_\theta$  and live in the same batch; the only additional cost over the pairwise baseline is one extra joint forward pass.

**Fusion vs. unimodal teacher.** A unimodal teacher  $g_m(x^{(m)})$ , e.g. SigLIP (Tschannen et al., 2025), Whisper, or BEATs (Chen et al., 2023a), is trained on the marginal  $p(x^{(m)})$  and therefore carries no information about cross-modal neighbors of  $x$ . By contrast,  $\mathbf{z}_{TVA}$  is computed on the joint input, so distillation into  $\mathbf{z}_A$  propagates  $\partial \text{sim}(\mathbf{z}_A, \text{sg}(\mathbf{z}_{TVA}))/\partial \theta_A$ , which encodes the  $T, V$  neighbors that an audio-only query will see at inference. This predicts two effects, both of which we confirm. Audio-anchored directions benefit most from  $\mathcal{L}_D$  (audio-related mean +3.73 vs. video-only +2.48, Table 5).

### 3.3 Tuple-InfoNCE Refinement ( $\mathcal{L}_T$ )

$\mathcal{L}_T$  complements pairwise alignment by directly optimizing the fused representation  $\mathbf{z}_{TVA}$ . Pairwise  $\mathcal{L}_A$  does not back-propagate through this joint vector, whereas  $\mathcal{L}_T$  contrasts two joint views

of a matched tuple against both ordinary in-batch negatives and tuples corrupted in exactly one modality. All embeddings below are  $\ell_2$ -normalized.

**Tuple-InfoNCE.** For tuple  $x_i = (T_i, V_i, A_i)$ , let  $a$  apply independent text, video, and audio augmentations, and let  $\tilde{x}_i$  be the single-slot corruption defined below. We compute

$$\mathbf{z}_i = \text{norm}(f_\theta(x_i)), \quad \mathbf{z}_i^+ = \text{norm}(f_\theta(a(x_i))),$$

$$\mathbf{h}_i = \text{sg}(\text{norm}(f_\theta(\tilde{x}_i))), \quad (3)$$

where  $\text{sg}$  stops gradients only through the corrupted tuple. For query batch  $Q$ , globally gathered key batch  $K$ , and local corrupted-tuple pool  $H$ , define

$$\mathcal{D}_i^{\text{key}}(Q, K) = \sum_{j \in \mathcal{G}} e^{\mathbf{q}_i^\top \mathbf{k}_j / \tau_T},$$

$$\mathcal{D}_i^{\text{hard}}(Q, H) = \sum_{j=1}^B e^{\mathbf{q}_i^\top \mathbf{h}_j / \tau_T},$$

$$\ell(Q, K, H) = -\frac{1}{B} \sum_i \log \frac{e^{\mathbf{q}_i^\top \mathbf{k}_{\pi(i)} / \tau_T}}{\mathcal{D}_i^{\text{key}} + \mathcal{D}_i^{\text{hard}}},$$

$$\mathcal{L}_T = \frac{1}{2} \{ \ell(Z, Z^+, H) + \ell(Z^+, Z, H) \}, \quad (4)$$

where  $\mathcal{G}$  indexes the differentially gathered global keys and  $\pi(i)$  is the matching global index. Both  $Z$  and  $Z^+$  receive gradients;  $H$  is detached and reused in both directions. Thus, unlike an average of unimodal similarities,  $\mathcal{L}_T$  directly supervises the fused embeddings.

**Modality-cycled hard negatives.** Each ordinary in-batch negative may differ from the anchor in all three modalities. We additionally construct one corrupted tuple per local sample by replacing exactly one slot. At step  $t$ , the slot is  $m_t = \{T, A, V\}[t \bmod 3]$ . Text uses a nonmatching caption from the same dataset type; audio and video use a local-batch derangement. Every query contrasts against the complete local pool  $H$ . This schedule exposes all three single-slot mismatch types without increasing the number of corrupted tuples per sample. Empirically,  $\mathcal{L}_D$  provides the dominant aggregate gain, while  $\mathcal{L}_T$  has a smaller, task-dependent effect (Tables 5 and 6).

### 3.4 Final Objective

*OmniRetriever* optimizes

$$\mathcal{L}_{\text{OmniRetriever}} = \lambda_D \mathcal{L}_D + \lambda_T \mathcal{L}_T + \lambda_A \mathcal{L}_A, \quad (5)$$

**Algorithm 1** *OmniRetriever* training step (batch  $\mathcal{B}$ , step  $t$ ).

- 
- 1: **Forward:**  $\mathbf{z}^{(T)}, \mathbf{z}^{(V)}, \mathbf{z}^{(A)} \leftarrow f_\theta$  (single-modal);  $Z \leftarrow \text{norm}(f_\theta(T, V, A))$
  - 2: **Pairwise**  $\mathcal{L}_A \leftarrow \sum_{(m, m') \in \{(T, V), (T, A), (V, A)\}} \mathcal{L}_{\text{NCE}}^{(m, m')}$
  - 3: **Distill**  $\mathcal{L}_D \leftarrow \frac{1}{3} \sum_{m \in \{T, V, A\}} \mathcal{L}_{\text{NCE}}^{(m, \text{sg}(Z))}$
  - 4: **Positive view:**  $Z^+ \leftarrow \text{norm}(f_\theta(a_T(T), a_V(V), a_A(A)))$
  - 5: **Cycled HN:**  $m_t \leftarrow \{T, A, V\}[t \bmod 3]$ ; replace slot  $m_t$  with a nonmatch to form  $\tilde{\mathcal{B}}$
  - 6: **Detached pool:**  $H \leftarrow \text{sg}(\text{norm}(f_\theta(\tilde{\mathcal{B}})))$
  - 7: **Tuple:**  $\mathcal{L}_T \leftarrow \frac{1}{2} [\ell(Z, Z^+, H) + \ell(Z^+, Z, H)]$
  - 8: **Update**  $\theta \leftarrow \theta - \eta \nabla_\theta (\lambda_D \mathcal{L}_D + \lambda_T \mathcal{L}_T + \lambda_A \mathcal{L}_A)$
- 

with  $(\lambda_D, \lambda_T, \lambda_A) = (1, 1, 1)$  chosen *a priori* (not tuned on *OmniRetriever-Bench*); Appendix D shows that the recipe is insensitive to the loss-weight ratio across a wide range. Algorithm 1 gives the per-step training loop. Relative to the pairwise baseline, the full method evaluates the original joint view  $Z$ , an augmented joint view  $Z^+$ , and a detached corrupted-tuple view  $H$ ; inference latency is reported in Appendix C.

## 4 Experiments

### 4.1 Setup

*OmniRetriever-7B* is an adapter fine-tune of the open-weights WAVE-7B (Tang et al., 2025) backbone, with LoRA adapters on the LLM trunk, an all-layer fusion head, and a BEATs adaptor; full architectural details are in Appendix A. The training data is a 1.5 M-triple subset sampled from four public video-text datasets, InternVid (Wang et al., 2024b), InternVid-FLT (Wang et al., 2024b), Panda-70M (Chen et al., 2024), and PVD (Bolya et al., 2025), together with a small in-house video corpus collected with consent. We restrict the pool to clips that contain text, video, and audio (Appendix E). We release model weights, evaluation code, and *OmniRetriever-Bench*. The training set is sample-identifier disjoint from every evaluation pool in this paper, and the third-party benchmarks (Clotho, SoundDescs, MSR-VTT, MSVD, DiDeMo, VATEX) are curated by other groups under different caption styles.

We compare *OmniRetriever-7B* (ours) against the frozen open-weights WAVE-7B (Tang et al., 2025) backbone and open Omni-Embed-Nemotron (Xu et al., 2025). On *OmniRetriever-Bench*, we additionally include a closed Gemini Embedding 2 configuration (Shanbhogue et al., 2026) as a descriptive API reference, not as

evidence for a training or routing claim. Our trainable configurations are reported as the mean over three seeds {42, 43, 44}; aggregate seed std is  $\leq 0.18$  R@1, which we treat as the noise floor. The PAIRWISE ablation ( $\mathcal{L}_A$  alone, the recipe used by recent unified embedders (Wei et al., 2024; Jiang et al., 2025; Zhang et al., 2025)) is reported in Table 5.

**Systems we do not compare against.** Qwen3-VL-Embedding (Li et al., 2026) has no audio path and cannot serve the ten audio-anchored directions. A head-to-head comparison is left to future work. The CLAP audio-text specialists (Table 2) have no video stream and cannot serve the six V-anchored directions.

### 4.2 Results on Standard Audio-Text Benchmarks

We evaluate *OmniRetriever-7B* zero-shot on two standard audio-text retrieval benchmarks, Clotho (Drossos et al., 2020) and SoundDescs (Koepke et al., 2023), and compare with the published audio-text specialist literature, contemporary open omni-modality embedders. Table 2 reports R@1; full R@ $k$  and NDCG@10 are in Table S7.

The CLAP-family specialists are audio-text dual-tower models trained on AudioCaps-style pairs without video exposure, and form the strongest single-task audio-retrieval band (19 to 21 R@1 on Clotho T→A). *OmniRetriever-7B* reaches 19.14 on Clotho T→A, within  $\sim 2$  R@1 of this band. On Clotho A→T, *OmniRetriever-7B* reaches 16.08, still below the specialist band (25 to 26.5) but well above prior open omni-modality systems on the same direction (ImageBind reports 6.0 T→A, LanguageBind 16.7 T→A; neither reports A→T). On SoundDescs T→A and A→T, *OmniRetriever-7B* reaches 25.00 and 20.70, the strongest open-system numbers in our literature search. Scaling the training corpus is expected to push *OmniRetriever-7B*’s A→T performance toward and potentially beyond the specialist band, since cross-modal supervision scales naturally with additional multimodal data.

### 4.3 Results on Standard Video-Text Benchmarks

We evaluate *OmniRetriever-7B* zero-shot on four standard video-text benchmarks (MSR-VTT (Xu et al., 2016), MSVD (Chen and Dolan, 2011),

Table 1: **Comparison with prior cross-modal retrieval benchmarks.** Each column under *Single-modal* or *Dual-modal* marks whether the benchmark supports a given retrieval direction-pair (a  $\checkmark$  indicates both directions of the pair are evaluated). *Total* is the resulting number of retrieval directions and *N* is the size of the standard test pool. *OmniRetriever-Bench* is the first public benchmark to cover all three AVT modalities in one pool and the first to evaluate dual-modal queries.

| Benchmark                            | Single-modal pairs    |                       |                       | Dual-modal pairs       |                        |                        | Total     | N            |
|--------------------------------------|-----------------------|-----------------------|-----------------------|------------------------|------------------------|------------------------|-----------|--------------|
|                                      | $T \leftrightarrow V$ | $T \leftrightarrow A$ | $V \leftrightarrow A$ | $T \leftrightarrow AV$ | $A \leftrightarrow TV$ | $V \leftrightarrow AT$ |           |              |
| <i>Video-text retrieval</i>          |                       |                       |                       |                        |                        |                        |           |              |
| MSR-VTT (Xu et al., 2016)            | ✓                     | —                     | —                     | —                      | —                      | —                      | 2         | 1,000        |
| MSVD (Chen and Dolan, 2011)          | ✓                     | —                     | —                     | —                      | —                      | —                      | 2         | 670          |
| DiDeMo (Anne Hendricks et al., 2017) | ✓                     | —                     | —                     | —                      | —                      | —                      | 2         | 1,004        |
| VATEX (Wang et al., 2019)            | ✓                     | —                     | —                     | —                      | —                      | —                      | 2         | 4,468        |
| <i>Audio-text retrieval</i>          |                       |                       |                       |                        |                        |                        |           |              |
| Clotho (Drossos et al., 2020)        | —                     | ✓                     | —                     | —                      | —                      | —                      | 2         | 1,045        |
| SoundDescs (Koepke et al., 2023)     | —                     | ✓                     | —                     | —                      | —                      | —                      | 2         | 1,000        |
| AudioCaps (Kim et al., 2019)         | —                     | ✓                     | —                     | —                      | —                      | —                      | 2         | 975          |
| <b>OmniRetriever-Bench (ours)</b>    | ✓                     | ✓                     | ✓                     | ✓                      | ✓                      | ✓                      | <b>12</b> | <b>3,782</b> |

Table 2: **Audio-text retrieval R@1 on Clotho and SoundDescs (zero-shot).** Specialised retrievers fine-tuned on each task are marked FT. Cited rows draw on the original papers’ published values. On Clotho  $T \rightarrow A$  *OmniRetriever-7B*’s 19.14 sits within  $\sim 2$  R@1 of the current zero-shot contrastive SOTA band (Cacophony, MGA-CLAP, M2D-CLAP at 20 to 21) without task-specific fine-tuning.

| Model                                        | Year | Setting | Clotho      |             | SoundDescs  |             |
|----------------------------------------------|------|---------|-------------|-------------|-------------|-------------|
|                                              |      |         | T→A         | A→T         | T→A         | A→T         |
| <i>Audio-text specialists</i>                |      |         |             |             |             |             |
| MMT (Koepke et al., 2023)                    | 2022 | FT      | 6.7         | 7.0         | 30.7        | 31.4        |
| CLAP-HTSAT (Deshmukh et al., 2023)           | 2023 | ZS      | 16.7        | 20.0        | —           | —           |
| LAION-CLAP fusion (Wu et al., 2023)          | 2023 | ZS      | 18.2        | 25.7        | —           | —           |
| WavCaps-HTSAT (Mei et al., 2024)             | 2023 | ZS      | 19.5        | 23.4        | —           | —           |
| WavCaps-CNN14 (Mei et al., 2024)             | 2023 | ZS      | <b>21.2</b> | 25.9        | —           | —           |
| FLAP fusion (Yeh et al., 2023)               | 2023 | ZS      | 20.3        | 25.5        | —           | —           |
| Cacophony (Zhu et al., 2024b)                | 2024 | ZS      | 20.2        | 26.5        | —           | —           |
| MGA-CLAP (Li et al., 2024)                   | 2024 | ZS      | 20.8        | <b>26.5</b> | —           | —           |
| M2D-CLAP (Niizumi et al., 2025)              | 2025 | ZS      | 20.1        | 25.0        | —           | —           |
| GLAP (Dinkel et al., 2026)                   | 2026 | ZS      | 19.4        | 21.8        | —           | —           |
| <i>Open-source omni-modality embedders</i>   |      |         |             |             |             |             |
| ImageBind (ViT-H) (Girdhar et al., 2023)     | 2023 | ZS      | 6.0         | —           | —           | —           |
| LanguageBind (CLIP-H/14) (Zhu et al., 2024a) | 2024 | ZS      | 16.7        | —           | —           | —           |
| Omni-Embed-Nemotron (Xu et al., 2025)        | 2025 | ZS      | 6.4         | 3.5         | 6.4         | 4.8         |
| <b>OmniRetriever-7B (ours)</b>               | 2026 | ZS      | <b>19.1</b> | <b>16.1</b> | <b>25.0</b> | <b>20.7</b> |

DiDeMo (Anne Hendricks et al., 2017), VATEX (Wang et al., 2019)). Table 3 reports R@1 on the evaluated pools; full R@k and NDCG@10 are in Table S6.

*OmniRetriever-7B* reaches 47.60/43.70 R@1 on MSR-VTT and 66.88/63.33 on MSVD. Because published systems can differ in preprocessing and evaluation pools, these values are reported without a data-parity claim. The controlled MSR-VTT loss ablation in Table 6 isolates the effect of the proposed objectives under one fixed pipeline.

**Why a unified AVT embedder.** Pairing a CLAP-style audio specialist with an InternVideo2 video specialist covers four of the twelve directions, but neither system produces embeddings for

the six dual-modal directions ( $T \leftrightarrow AV$ , etc.) in one shared index, and neither can anchor an audio query against text grounded in shared video context. *OmniRetriever* targets the setting in which a single shared ( $T, V, A$ ) index is required.

*OmniRetriever-7B* serves audio-text, video-text, and the six dual-modal directions with one shared encoder and index.

#### 4.4 *OmniRetriever-Bench*: A 12-Direction AVT Probe

The benchmarks above evaluate single-modal directions only and do not stress-test the joint representation  $\mathbf{z}_{TVA}$  that *OmniRetriever* is built around. A practical AVT retriever also needs to serve six dual-modal queries ( $T \leftrightarrow AV$ ,  $A \leftrightarrow TV$ ,

Table 3: **Zero-shot video–text retrieval with *OmniRetriever-7B***. R@1 is reported on each evaluated pool;  $N$  is the number of query–target pairs.

| Benchmark | $N$   | $T \rightarrow AV$ | $AV \rightarrow T$ |
|-----------|-------|--------------------|--------------------|
| MSR-VTT   | 1,000 | 47.60              | 43.70              |
| MSVD      | 480   | 66.88              | 63.33              |
| DiDeMo    | 315   | 46.03              | 45.08              |
| VATEX     | 2,494 | 57.98              | 54.97              |

$V \leftrightarrow AT$ ) that compose two modalities on one side. No public suite covers these directions: standard video–text and audio–text benchmarks each evaluate two single-modal directions, never both video and audio in the same pool. [Table 1](#) contrasts *OmniRetriever-Bench* with these prior benchmarks across modality coverage, number of retrieval directions, and evaluation-pool size.

*OmniRetriever-Bench* contains 3,782 held-out aligned  $(T, V, A)$  triples, scored against the same gallery in every direction; the metric is Recall@1. Triples are curated from in-house video sources that are disjoint from the training data at the sample-identifier level. **All 3,782 captions are reviewed and corrected by trained human annotators:** twenty trained in-house annotators revised Gemini 3.0 Pro drafts against both the video and audio, after which a different annotator performed a second review. The second-pass acceptance rate exceeded 90%. We release source identifiers, clip intervals, and captions under a research-use license, but do not redistribute the underlying media. Full annotation guidelines, the distribution mismatch with the training corpus, and a caption-paraphrase robustness check are in [Appendices E, F and L](#).

[Table 4](#) reports R@1 on all 12 directions of *OmniRetriever-Bench*. *OmniRetriever-7B* reaches AVG-all **34.84**, +9.52 over the frozen WAVE-7B backbone and +8.03 over Omni-Embed-Nemotron. These open-system gaps exceed the seed noise floor ( $\leq 0.18$  R@1 std). The closed-API row is included only as an external reference on the same benchmark pool and is not used for a data-parity or causal claim.

#### 4.5 Ablation and Analysis

[Table 5](#) reports the per-direction stair-step ablation against the frozen WAVE-7B base on *OmniRetriever-Bench*, isolating the contribution of each loss in the *OmniRetriever* objective.

**Pairwise baseline.** The PAIRWISE baseline ( $\mathcal{L}_A$  only, on our training corpus) lifts AVG-all from

25.32 to 31.08 (+5.76). The gain is unevenly distributed:  $T \leftrightarrow V$  and most dual-modal directions improve by 5 to 9 R@1, but audio-anchored R@1 stays low ( $a \rightarrow t$  9.00,  $a \rightarrow tv$  17.00). This observation does not isolate whether the remaining gap comes from data, routing, or capacity; it motivates the matched joint-objective ablations below.

**Fusion-as-teacher distillation ( $\mathcal{L}_D$ ).** Adding  $\mathcal{L}_D$  lifts AVG-all by +3.52 (31.08  $\rightarrow$  34.60), the dominant single-loss contribution. The gain spans both regimes (video-only +2.48, audio-related +3.73; [Table 5](#)), consistent with the joint forward providing cross-modal context to the single-modal students.

**Tuple-InfoNCE refinement ( $\mathcal{L}_T$ ).** Adding  $\mathcal{L}_T$  on top of  $\mathcal{L}_D + \mathcal{L}_A$  lifts AVG-all by +0.24 to **34.84**. The aggregate gain is small but stable across seeds: the per-seed  $\Delta_{\mathcal{L}_T}$  is  $\{+0.24, +0.21, +0.27\}$  on  $\{42, 43, 44\}$  ([Table S9](#)), well above the aggregate seed std ( $\leq 0.07$ ). The per-direction breakdown ([Table 5](#)) shows that  $\mathcal{L}_T$  redistributes capacity: it adds +2.03 to +2.86 R@1 on the four  $A \leftrightarrow V$  routes (above per-direction noise of  $\pm 0.5$  to  $\pm 0.8$ ) while losing  $\sim 2$  R@1 on text-anchored dual-modal routes. We keep  $\mathcal{L}_T$  in the released model as a complementary term. Controlled external ablations show a task-dependent contribution: +2.50 R@1 on MSR-VTT  $T \rightarrow V$  and +0.91 on Clotho  $T \rightarrow A$  ([Table 6](#)). Alternative joint-view, joint-only, regression, and EMA controls are reported in [Table 7](#).

**Discussion.** Pairwise InfoNCE optimizes three pair softmaxes independently, so the joint  $(T, V, A)$  vector inherits whichever pair carries the largest cosine gap (in practice  $T-V$ ) and leaves the audio axis slack.  $\mathcal{L}_D$  supplies the cross-modal teacher that closes the audio gap at the single-modal level.  $\mathcal{L}_T$  then redistributes the remaining joint-level capacity toward the  $A \leftrightarrow V$  axis.

Table 4: **Recall@1 on the 12 directions of OmniRetriever-Bench.** **Bold:** best per direction. The **ours** row is shaded. The pairwise-only baseline (PAIRWISE,  $\mathcal{L}_A$  alone) is reported in Table 5.

| System                                      | Single-modal      |                   |                   |                   |                   |                   | Dual-modal         |                    |                    |                    |                    |                    | Averages     |              |              |
|---------------------------------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------|--------------|--------------|
|                                             | $t \rightarrow v$ | $v \rightarrow t$ | $t \rightarrow a$ | $a \rightarrow t$ | $v \rightarrow a$ | $a \rightarrow v$ | $t \rightarrow av$ | $av \rightarrow t$ | $a \rightarrow tv$ | $tv \rightarrow a$ | $v \rightarrow at$ | $at \rightarrow v$ | single       | dual         | all          |
| <i>External closed-API reference</i>        |                   |                   |                   |                   |                   |                   |                    |                    |                    |                    |                    |                    |              |              |              |
| Gemini Embedding 2 (Shanhogue et al., 2026) | <b>55.13</b>      | <b>53.83</b>      | 12.61             | 1.48              | 13.80             | 15.76             | <b>55.45</b>       | <b>50.16</b>       | 6.00               | 16.79              | <b>54.97</b>       | <b>61.45</b>       | 25.44        | 40.80        | 33.12        |
| <i>Open-source omni-modality embedders</i>  |                   |                   |                   |                   |                   |                   |                    |                    |                    |                    |                    |                    |              |              |              |
| WAVE-7B (Tang et al., 2025)                 | 41.27             | 34.45             | 8.67              | 3.38              | 14.89             | 12.93             | 44.95              | 42.68              | 2.27               | 15.65              | 38.74              | 43.92              | 19.27        | 31.37        | 25.32        |
| Omni-Embed-Nemotron (Xu et al., 2025)       | 43.55             | 39.74             | 12.19             | 8.83              | 14.07             | 12.35             | 36.12              | 34.53              | 14.75              | 16.63              | 41.30              | 47.67              | 21.79        | 31.84        | 26.81        |
| <b>OmniRetriever (ours)</b>                 | 48.89             | 46.30             | <b>14.20</b>      | <b>11.92</b>      | <b>25.46</b>      | <b>24.99</b>      | 45.37              | 44.47              | <b>23.45</b>       | <b>26.07</b>       | 52.49              | 54.47              | <b>28.63</b> | <b>41.05</b> | <b>34.84</b> |

Table 5: **Per-direction stair-step ablation on OmniRetriever-Bench.** Each  $\Delta$  is the increment over the preceding stage; the columns sum to the final OmniRetriever-7B score.

| Direction          | WAVE-7B | $\Delta_{\text{CL}}$ | $\Delta_{\mathcal{L}_D}$ | $\Delta_{\mathcal{L}_T}$ | OmniRetriever-7B |
|--------------------|---------|----------------------|--------------------------|--------------------------|------------------|
| $t \rightarrow v$  | 41.27   | +4.73                | +3.13                    | -0.24                    | 48.89            |
| $v \rightarrow t$  | 34.45   | +10.55               | +1.83                    | -0.53                    | 46.30            |
| $t \rightarrow a$  | 8.67    | +3.33                | +2.46                    | -0.26                    | 14.20            |
| $a \rightarrow t$  | 3.38    | +5.62                | +2.42                    | +0.50                    | 11.92            |
| $v \rightarrow a$  | 14.89   | +6.11                | +2.43                    | +2.03                    | 25.46            |
| $a \rightarrow v$  | 12.93   | +3.07                | +6.16                    | +2.83                    | 24.99            |
| $t \rightarrow av$ | 44.95   | +1.05                | +1.70                    | -2.33                    | 45.37            |
| $av \rightarrow t$ | 42.68   | +2.32                | +1.35                    | -1.88                    | 44.47            |
| $a \rightarrow tv$ | 2.27    | +14.73               | +8.28                    | -1.83                    | 23.45            |
| $tv \rightarrow a$ | 15.65   | +7.35                | +3.84                    | -0.77                    | 26.07            |
| $v \rightarrow at$ | 38.74   | +6.26                | +4.63                    | +2.86                    | 52.49            |
| $at \rightarrow v$ | 43.92   | +4.08                | +3.98                    | +2.49                    | 54.47            |
| <b>AVG all</b>     | 25.32   | +5.76                | <b>+3.52</b>             | +0.24                    | <b>34.84</b>     |

#### 4.6 Controlled Validation

We test whether the loss attribution survives changes in the evaluation pool, objective design, and backbone while holding the remaining factors fixed within each comparison.

**External benchmarks.** Under the matched WAVE-7B pipeline,  $\mathcal{L}_D$  adds +3.00 R@1 over pairwise alignment on MSR-VTT T $\rightarrow$ V and +1.80 on Clotho T $\rightarrow$ A;  $\mathcal{L}_T$  adds a further +2.50 and +0.91, respectively. The joint losses therefore remain useful outside OmniRetriever-Bench, with a task-dependent effect.

Table 6: **Controlled loss ablations on external benchmarks** (R@1). All fine-tuned rows hold the backbone, data, schedule, and evaluation code fixed.

| Dataset / direction       | Frozen | $\mathcal{L}_A$ | $\mathcal{L}_A + \mathcal{L}_D$ | Full         |
|---------------------------|--------|-----------------|---------------------------------|--------------|
| MSR-VTT T $\rightarrow$ V | 36.90  | 42.10           | 45.10                           | <b>47.60</b> |
| Clotho T $\rightarrow$ A  | 14.08  | 16.43           | 18.23                           | <b>19.14</b> |

**Alternative objectives.** Direct joint-view contrastive training alone remains at the pairwise baseline, while joint-only training fails to preserve a useful unimodal space. Normalized-MSE distillation trails the contrastive teacher by 0.95 R@1. An EMA teacher is within 0.05 R@1 of the same-step fused teacher but requires additional teacher state.

Table 7: **Alternative objective controls on OmniRetriever-Bench** (AVG-all R@1). Backbone, training data, schedule, and evaluation are held fixed.

| Objective                                               | AVG-all      |
|---------------------------------------------------------|--------------|
| Pairwise $\mathcal{L}_A$                                | 31.08        |
| $\mathcal{L}_A$ + direct joint-view contrastive control | 31.12        |
| Joint supervision only                                  | 20.67        |
| Replace $\mathcal{L}_D$ with normalized MSE             | 33.89        |
| EMA teacher for $\mathcal{L}_D$                         | 34.79        |
| <b>Full method</b>                                      | <b>34.84</b> |

**Cross-backbone transfer.** On the architecturally different Omni-Embed-Nemotron-3B,  $\mathcal{L}_D$  again supplies the dominant added gain (+2.43 over pairwise), and  $\mathcal{L}_T$  contributes a smaller +0.33. This reproduces the WAVE-7B attribution pattern under a short 3,000-step transfer run; full setup details are in Appendix K.

Table 8: **Cross-backbone validation on Omni-Embed-Nemotron-3B** (3,000 LoRA steps each, OmniRetriever-Bench AVG-all R@1;  $\Delta$  vs. the frozen base).

| Variant                                               | AVG all      | $\Delta$ vs. base |
|-------------------------------------------------------|--------------|-------------------|
| Nemotron-3B (frozen)                                  | 26.81        | —                 |
| PAIRWISE ( $\mathcal{L}_A$ only)                      | 29.42        | +2.61             |
| $\mathcal{L}_D + \mathcal{L}_A$ (no $\mathcal{L}_T$ ) | 31.85        | +5.04             |
| <b>Full OmniRetriever</b>                             | <b>32.18</b> | <b>+5.37</b>      |

#### 4.7 Sensitivity, Efficiency, and Failure Modes

**Loss-weight sensitivity.** The released model uses uniform weights  $(\lambda_T, \lambda_D, \lambda_A) = (1, 1, 1)$ . A deliberately skewed (50, 10, 1) setting remains above the pairwise baseline (31.08 AVG-all) and changes AVG-all by only -0.71 R@1 (Table 9). The uniform setting provides the stronger single- and dual-modal averages; the full per-direction comparison is in Appendix D.

Table 9: **Loss-weight sensitivity on *OmniRetriever-Bench* R@1.**  $\Delta$  compares (50, 10, 1) with the released uniform setting.

| Metric     | (1, 1, 1)    | (50, 10, 1) | $\Delta$ |
|------------|--------------|-------------|----------|
| AVG single | <b>28.63</b> | 28.09       | -0.54    |
| AVG dual   | <b>41.05</b> | 40.17       | -0.88    |
| AVG all    | <b>34.84</b> | 34.13       | -0.71    |

**Joint-space geometry.** We measure the mean triple-cosine score between the three modalities for matched triples and for mismatched triples in *OmniRetriever-Bench*. Although InfoNCE disperses absolute cosine values, the matched-minus-mismatched gap grows from 0.060 in the frozen base to 0.078 after pairwise training and 0.080 with the full method (Table 10).

Table 10: **Mean triple-cosine geometry on *OmniRetriever-Bench*** ( $N = 3,782$ ). The gap is matched minus mismatched.

| Variant                          | matched      | mismatched   | gap          |
|----------------------------------|--------------|--------------|--------------|
| WAVE-7B (frozen)                 | 0.495        | 0.435        | 0.060        |
| PAIRWISE ( $\mathcal{L}_A$ only) | 0.375        | 0.297        | 0.078        |
| <b>Full <i>OmniRetriever</i></b> | <b>0.322</b> | <b>0.243</b> | <b>0.080</b> |

**Inference cost.** Table 11 reports full-stack single-query latency at bf16 and batch size 1, including the modality towers, LLM trunk, and fusion head. Joint AV is the most expensive path at 125.1 ms. Peak memory over all paths is 20.13 GB, of which model loading accounts for 17.92 GB.

Table 11: **End-to-end inference latency of *OmniRetriever-7B*.** Mean $\pm$ std over 10 iterations after 3 warmups; one RTX PRO 6000 Blackwell Max-Q GPU.

| Inference path | Latency (ms)    |
|----------------|-----------------|
| Text-only      | 20.9 $\pm$ 0.1  |
| Video-only     | 105.6 $\pm$ 0.2 |
| Audio-only     | 34.6 $\pm$ 0.1  |
| Joint AV       | 125.1 $\pm$ 0.1 |

**Failure modes.** Among 60 sampled queries (5 per direction) whose gold target falls outside the top 10, 74% place it at ranks 11–50. The highest-ranked alternatives are often semantic near-ties, such as dialogue clips differing mainly in speaker identity or product B-roll differing in object sub-category. This suggests that many residual errors require fine-grained instance discrimination rather than broader cross-modal alignment.

## 5 Conclusion

A unified AVT encoder can compute a joint embedding  $\mathbf{z}_{TVA}$  on every  $(T, V, A)$  forward, but

pairwise InfoNCE leaves it unsupervised. We turn  $\mathbf{z}_{TVA}$  into a training signal:  $\mathcal{L}_D$  distills its stop-gradient copy into the single-modal embeddings, and  $\mathcal{L}_T$  supervises it directly against a tuple grid with modality-cycled hard negatives. Both losses reuse the same backbone parameters without an external teacher and apply to any unified retriever whose forward produces a joint embedding. A cross-backbone replication reproduces the dominant  $\mathcal{L}_D$  effect. *OmniRetriever-7B* reaches the zero-shot audio–text specialist band on Clotho  $T \rightarrow A$ , reduces the open omni-modality gap on  $A \rightarrow T$ , and reaches 34.84 AVG-all R@1 on *OmniRetriever-Bench*. To probe the joint representation that pairwise training cannot expose, we release *OmniRetriever-Bench*, the first AVT benchmark scoring 3,782 human-corrected triples on a shared gallery across all 12 single- and dual-modal directions. Model weights, training code, and the benchmark are released.

## Limitations

Comparisons with closed systems and published results are contextual because their training data, preprocessing, and service implementations are not controlled. We therefore do not make dataparity or causal claims from those comparisons. A unified  $(T, V, A)$  embedder at our training scale is also not expected to match dedicated single-task encoders on every benchmark. Beyond accuracy, our analyses do not study embedding compression beyond the post-hoc int8/binary baselines reported in Appendix G; a compression-aware training recipe is left to future work.

## Ethical Considerations

The training corpus combines four public video–text datasets (InternVid, InternVid-FLT, Panda-70M, PVD) with a small in-house corpus collected with consent. We do not redistribute the assembled training subset. *OmniRetriever-Bench* is built from sources held out from training and is released following the standard video-benchmark protocol: source identifiers, clip intervals, and captions are released, while the underlying media is not redistributed. We release *OmniRetriever-7B* and *OmniRetriever-Bench* for research on cross-modal retrieval. The research-use license accompanying the release prohibits deployment in surveillance applications affecting natural persons.

## References

- Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. [VATT: Transformers for multimodal self-supervised learning from raw video, audio and text](#). In *Advances in Neural Information Processing Systems*, volume 34.
- Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.
- Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zhohus, et al. 2025. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*.
- Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. 2022. [MultiMAE: Multi-modal multi-task masked autoencoders](#). In *Computer Vision – ECCV 2022*, volume 13697 of *Lecture Notes in Computer Science*, pages 348–367. Springer.
- Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. [LLM2Vec: Large language models are secretly powerful text encoders](#). In *First Conference on Language Modeling*.
- Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. 2025. [Perception encoder: The best visual embeddings are not at the output of the network](#). *arXiv preprint arXiv:2504.13181*.
- David Chen and William B Dolan. 2011. [Collecting highly parallel data for paraphrase evaluation](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA. Association for Computational Linguistics.
- Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xi-angzhan Yu, and Furu Wei. 2023a. [BEATs: Audio pre-training with acoustic tokenizers](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5178–5193. PMLR.
- Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023b. [VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-Wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. 2024. [Panda-70M: Captioning 70m videos with multiple cross-modality teachers](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13320–13331.
- Soham Deshmukh, Benjamin Elizalde, and Huaming Wang. 2023. [Audio retrieval with WavText5K and CLAP training](#). In *Interspeech 2023*, pages 2948–2952.
- Heinrich Dinkel, Zhiyong Yan, Tianzi Wang, Yongqing Wang, Xingwei Sun, Yadong Niu, Jizhong Liu, Gang Li, Junbo Zhang, and Jian Luan. 2026. [GLAP: General contrastive audio-text pretraining across domains and languages](#). In *ICASSP 2026 – 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 14737–14741.
- Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.
- Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. 2024. [Data filtering networks](#). In *The Twelfth International Conference on Learning Representations*.
- Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G Turrissi da Costa, Louis Béthune, Zhe Gan, et al. 2025. Multimodal autoregressive pre-training of large vision encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9654.
- Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. 2022. [Multi-modal masked autoencoders learn transferable representations](#). *arXiv preprint arXiv:2205.14204*.
- Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Man-nat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190.
- Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. [AudioCLIP: Extending CLIP to image, text and audio](#). In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

- Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. 2025. [VLM2Vec: Training vision-language models for massive multi-modal embedding tasks](#). In *The Thirteenth International Conference on Learning Representations*.
- Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. [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, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.
- A. Sophia Koepke, Andreea-Maria Oncescu, João F. Henriques, Zeynep Akata, and Samuel Albanie. 2023. [Audio retrieval with natural language queries: A benchmark study](#). *IEEE Transactions on Multimedia*, 25:2675–2685.
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. [NV-Embed: Improved techniques for training LLMs as generalist embedding models](#). In *The Thirteenth International Conference on Learning Representations*.
- Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibao Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. 2026. [Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking](#). *arXiv preprint arXiv:2601.04720*.
- Yiming Li, Zhifang Guo, Xiangdong Wang, and Hong Liu. 2024. [Advancing multi-grained alignment for contrastive language-audio pre-training](#). In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7356–7365.
- Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. [MM-Embed: Universal multimodal retrieval with multimodal LLMs](#). In *The Thirteenth International Conference on Learning Representations*.
- Yunze Liu, Qingnan Fan, Shanghang Zhang, Hao Dong, Thomas A. Funkhouser, and Li Yi. 2021. [Contrastive multimodal fusion with TupleInfoNCE](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 754–763.
- Yunze Liu, Li Yi, Shanghang Zhang, Qingnan Fan, Thomas A. Funkhouser, and Hao Dong. 2020. [P4Contrast: Contrastive learning with pairs of point-pixel pairs for RGB-D scene understanding](#). *arXiv preprint arXiv:2012.13089*.
- Alina Marcu, Mihai Pirvu, Dragos Costea, Emanuela Haller, Emil Slusanschi, Ahmed Nabil Belbachir, Rahul Sukthankar, and Marius Leordeanu. 2023. [Self-supervised hypergraphs for learning multiple world interpretations](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 983–992.
- Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuejian Zou, and Wenwu Wang. 2024. [Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354.
- Daisuke Niizumi, Daiki Takeuchi, Masahiro Yasuda, Binh Thien Nguyen, Yasunori Ohishi, and Noboru Harada. 2025. [M2D-CLAP: Exploring general-purpose audio-language representations beyond CLAP](#). *IEEE Access*, 13:163313–163330.
- Mihai Pirvu, Alina Marcu, Maria Alexandra Dobrescu, Ahmed Nabil Belbachir, and Marius Leordeanu. 2023. [Multi-task hypergraphs for semi-supervised learning using earth observations](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3404–3414.
- Mihai-Cristian Pirvu and Marius Leordeanu. 2025. [Probabilistic hyper-graphs using multiple randomly masked autoencoders for semi-supervised multi-modal multi-task learning](#). *arXiv preprint arXiv:2510.10068*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. [Learning transferable visual models from natural language supervision](#). In *International conference on machine learning*, pages 8748–8763. PmlR.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. [Laion-5b: An open large-scale dataset for training next generation image-text models](#). *Advances in neural information processing systems*, 35:25278–25294.
- Madhuri Shanbhogue, Zhe Li, Shanfeng Zhang, Gustavo Hernández Ábrego, Shih-Cheng Huang, Aashi Jain, Daniel Salz, Sonam Goenka, Chaitra Hegde, Ji Ma, et al. 2026. [Gemini Embedding 2: A native multimodal embedding model from Gemini](#). *arXiv preprint arXiv:2605.27295*.
- Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. 2025. [Dinov3](#). *arXiv preprint arXiv:2508.10104*.
- Changli Tang, Qinfan Xiao, Ke Mei, Tianyi Wang, Fengyun Rao, and Chao Zhang. 2025. [Wave: Learning unified & versatile audio-visual embeddings with multimodal llm](#). *arXiv preprint arXiv:2509.21990*.
- Michael Tschanen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer,

- Ye Xia, Basil Mustafa, et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. [Improving text embeddings with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.
- Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuanfang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591.
- Yi Wang, Yanan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. 2024b. [InternVid: A large-scale video-text dataset for multimodal understanding and generation](#). In *The Twelfth International Conference on Learning Representations*.
- Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pages 387–404. Springer.
- Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2024. [Demystifying CLIP data](#). In *The Twelfth International Conference on Learning Representations*.
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.
- Mengyao Xu, Wenfei Zhou, Yauhen Babakhin, Gabriel Moreira, Ronay Ak, Radek Osmulski, Bo Liu, Even Oldridge, and Benedikt Schifferer. 2025. Omni-embed-nemotron: A unified multimodal retrieval model for text, image, audio, and video. *arXiv preprint arXiv:2510.03458*.
- Ching-Feng Yeh, Po-Yao Huang, Vasu Sharma, Shang-Wen Li, and Gargi Gosh. 2023. [FLAP: Fast language-audio pre-training](#). In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.
- Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025. [Bridging modalities: Improving universal multimodal retrieval by multimodal large language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9274–9285.
- Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiayi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024a. [LanguageBind: Extending video-language pretraining to N-modality by language-based semantic alignment](#). In *The Twelfth International Conference on Learning Representations*.
- Ge Zhu, Jordan Darefsky, and Zhiyao Duan. 2024b. [Cacophony: An improved contrastive audio-text model](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4867–4879.

## Appendix Overview

- [Appendices A and B](#): backbone, LoRA configuration, and full training hyper-parameters.
- [Appendix C](#): measurement details for the main-paper inference-latency table.
- [Appendix D](#): full per-direction sensitivity of *OmniRetriever* to the loss weights ( $\lambda_T$ ,  $\lambda_D$ ,  $\lambda_A$ ).
- [Appendix E](#): training-corpus description and caption-length statistics.
- [Appendices F and H](#): *OmniRetriever-Bench* construction, licensing, and test-clip duration statistics.
- [Appendix G](#): post-hoc compression analysis (uniform dimension downsampling, int8 and binary quantization, front-dim truncation).
- [Appendix I](#): per- $k$  R@1/5/10 and NDCG@10 on all six external benchmarks.
- [Appendix J](#): definition and mismatch construction for the main-paper triple-cosine analysis.
- [Appendix K](#): training details for the cross-backbone replication on Omni-Embed-Nemotron-3B.
- [Appendix L](#): caption-paraphrase robustness check on *OmniRetriever-Bench*.
- [Appendix M](#): per-seed  $\mathcal{L}_T$  ablation across three seeds.

## A Backbone and LoRA Fine-Tuning Details

We take WAVE-7B (Tang et al., 2025), an open-weights extension of Qwen2.5-Omni with a hierarchical visual fusion and a dual audio encoder, as our *baseline backbone*. WAVE-7B is the strongest open starting point on tri-modal tasks but, like all prior open systems, relies on pairwise contrastive losses and lags closed competitors on audio-involved retrieval; *OmniRetriever* retrofits the joint-objective recipe on top of it. The choice of WAVE-7B (vs. alternative AVT backbones) is incidental to the *OmniRetriever* contribution:  $\mathcal{L}_D$  and  $\mathcal{L}_T$  act on the backbone’s joint and single-modal embeddings, not on its internal architecture, and would transfer to any AVT model that exposes a fusion path. We *freeze* the entire WAVE-7B backbone, namely the LLM trunk, the visual tower (SigLIP + merger), the audio tower (Whisper), and the BEATs encoder transformer (all adopted from WAVE without modification), and tune three small surfaces: (i) LoRA adapters ( $r=16, \alpha=32$ ) inserted into the  $q, k, v$  projections of *every* LLM layer (28 layers  $\times$  3 = 84 matrices, 6.88 M params); (ii) the all-layer fusion head  $\text{Linear}(28d \rightarrow d) \rightarrow \text{GELU} \rightarrow \text{Linear}(d \rightarrow d)$  with  $d=3584$ , 372.51 M params; (iii) the BEATs adaptor on top of the frozen BEATs encoder (Layer-Norm + projector, 15.61 M params). The fusion head consumes the concatenation of the last-token hidden states of all 28 LLM layers and projects them into the shared  $\mathbb{R}^{3584}$  embedding space; sub-modal embeddings reuse the same head. Total trainable parameters:  $\approx 395.0$  M (4.20 % of the 9.41 B backbone). Training uses bf16 + DeepSpeed ZeRO-0, batch size 64, LR  $1 \times 10^{-5}$  cosine, 1 epoch over the training corpus, on 4 $\times$ NVIDIA RTX PRO 6000 (Blackwell architecture, Max-Q variant, 96 GB). Total wall-clock:  $\approx 109$  h (35.6 s/step over 11,192 steps).

## B Hyper-Parameters and Reproducibility

Table S1 lists every hyper-parameter that affects the reported numbers. We run all reported numbers under three independent seeds {42, 43, 44} and report the mean; aggregate seed std on *OmniRetriever-Bench* AVG-all is  $\leq 0.18$  R@1.

## C Inference Measurement Details

Table 11 measures full-stack single-query latency on one RTX PRO 6000 Blackwell Max-Q GPU at bf16 and batch size 1, averaged over 10 timed iterations after 3 warmups. The forward pass includes the modality towers, LM trunk with all-layer hidden states retained, and the  $28d \rightarrow d \rightarrow d$  fusion head. Text uses 24 tokens; video uses 8 frames at 224 px ( $\sim 268$  LM tokens); audio uses 8 s of 16 kHz mono; joint AV concatenates video, audio, and prompt tokens ( $\sim 470$  total). Cosine search and text-side preprocessing are excluded.

## D Loss-Weight Sensitivity

The released *OmniRetriever-7B* uses uniform loss weights  $(\lambda_T, \lambda_D, \lambda_A) = (1, 1, 1)$ . To test sensitivity to this choice, we re-train *OmniRetriever* from the same WAVE-7B initialisation under the skewed setting  $(\lambda_T, \lambda_D, \lambda_A) = (50, 10, 1)$ , which scales the Tuple-InfoNCE loss by  $50\times$  and the distillation loss by  $10\times$  relative to pairwise alignment. Table S2 reports both configurations on the 3,782-sample *OmniRetriever-Bench*.

**Discussion.** Both configurations improve over the pairwise-CL ablation (31.08), so the method’s central claim is unaffected by the specific weighting. The heavily-anchor-weighted (50, 10, 1) setting slightly improves audio-anchored directions ( $a \rightarrow v +0.74$ ,  $a \rightarrow tv +1.19$ ) at the cost of text-anchored ones ( $t \rightarrow av -1.95$ ,  $av \rightarrow t -2.27$ ): putting more loss mass on the fusion anchor pulls the audio sub-encoder closer to it but de-emphasizes the pairwise alignment that benefits text-anchored composition. The uniform setting is a Pareto-friendly default. Per-pair temperature tuning or curriculum schedules over  $(\lambda_T, \lambda_D, \lambda_A)$  are left as future work.

## E Training Corpus Description

**Sources and scale.** The training corpus contains approximately 1.5 M ( $T, V, A$ ) triples sampled from four public video–text datasets, InternVid (Wang et al., 2024b), InternVid-FLT (Wang et al., 2024b), Panda-70M (Chen et al., 2024), and PVD (Bolya et al., 2025), plus a small in-house video corpus collected with consent. We restrict the pool to clips that contain all three modalities: a text caption, video frames, and an audible audio track. We do not redistribute the assembled training subset; the public sources can be downloaded

Table S1: **Full hyper-parameters for *OmniRetriever* training. Tunable surfaces:** LoRA on the q,k,v projections of all 28 LLM layers (84 matrices) plus, at full precision, the all-layer fusion head and the BEATs adaptor (Layer-Norm + projector). **Frozen:** LLM trunk weights, vision tower (incl. merger), Whisper audio tower, and the BEATs encoder backbone. **Trainable budget:**  $\approx 395.0$  M total (4.20% of the 9.41 B backbone), distributed as 6.88 M (LoRA), 372.51 M (all-layer fusion head: Linear( $28d \rightarrow d$ )+GELU+Linear( $d \rightarrow d$ ) with  $d = 3584$ ), and 15.61 M (BEATs adaptor). **Loss settings:** single shared temperature  $\tau$  (no per-pair  $\tau_{mm'}$ ); hard negatives are deterministically cycled {T-shuffle, A-shuffle, V-shuffle} with period 3, tightening the joint cluster on every modality direction (Section 3.3 in main paper).

| Hyper-parameter                                    | Value                                      |
|----------------------------------------------------|--------------------------------------------|
| Backbone (baseline)                                | WAVE-7B (Tang et al., 2025)                |
| LoRA rank $r$ / $\alpha$ / dropout                 | 16 / 32 / 0.05                             |
| Embedding dim $d$                                  | 3584                                       |
| Text max length                                    | 2048 tokens                                |
| Video frames / Audio duration                      | 8 / 8 s @ 16 kHz                           |
| Audio encoders / Vision teacher                    | Whisper-Large-v3 + BEATs / SigLIP-2-SO400m |
| Optimiser                                          | AdamW ( $\beta_1=0.9, \beta_2=0.95$ )      |
| Learning rate / Warm-up / WD                       | $10^{-5}$ cosine / 3% / 0.01               |
| Batch size                                         | 64 (8/GPU $\times$ 4 GPUs $\times$ 2 acc)  |
| Epochs / Precision / DeepSpeed                     | 1 / bf16 / ZeRO-0                          |
| Tuple-InfoNCE temperature $\tau$                   | 0.01                                       |
| Loss weights ( $\lambda_T, \lambda_D, \lambda_A$ ) | (1.0, 1.0, 1.0)                            |
| Hardware                                           | 4 $\times$ RTX PRO 6000 Blackwell-MaxQ     |
| Total wall-clock (11,192 steps)                    | $\approx 109$ h (35.6 s/step)              |

from their original releases. *OmniRetriever-Bench* is built from sources held out from training, with sample-identifier disjointness enforced.

**Caption sources.** For triples drawn from the public datasets, we keep the captions provided by the original release. For the in-house portion, captions are produced by an automatic multimodal captioning pipeline with a held-out coherence filter on  $(T, V, A)$  alignment. *OmniRetriever-Bench* captions, in contrast, are human-corrected: each caption begins as a Gemini 3.0 Pro draft and is then verified against the video and audio and rewritten where necessary by a trained annotator. The resulting *OmniRetriever-Bench* captions therefore reflect a distinct human-validated distribution from any caption in the training corpus. We additionally test sensitivity to caption surface form with the paraphrase analysis in Appendix L.

**Human review and quality control.** Twenty trained in-house annotators, compensated through their salaried positions, processed the 3,782 benchmark captions. One annotator corrected each draft against the visible and audible content, and a different annotator conducted a second review. More than 90% were accepted on that second pass. Exact-caption and near-duplicate screening was performed manually before inclusion.

**Caption length distribution.** Table S3 reports caption-word quantiles measured on the training corpus. Captions are dense and uniform, with mean 32.0 words, median 30, and 99<sup>th</sup> percentile 76.

## F *OmniRetriever-Bench*: Construction and License

The *OmniRetriever-Bench* evaluation pool contains 3,782 held-out  $(T, V, A)$  triples; the same pool serves as queries *and* gallery for all twelve retrieval directions. *OmniRetriever-Bench* is released for research use under a custom license: caption text, public source identifiers, and retrieval indices are redistributed, while the underlying media is not. *OmniRetriever-Bench* triples come from internal video sources held out from the training subset, with sample-identifier disjointness from the training data enforced. Figure S1 visualises the 12 valid query $\rightarrow$ target cells of the  $6 \times 6$  source/target matrix.

**Distribution.** Following the protocol of prior video-language benchmarks (Wang et al., 2019; Xu et al., 2016; Chen and Dolan, 2011), we redistribute the *OmniRetriever-Bench* evaluation pipeline, the held-out caption text, source video identifiers, and clip time intervals; the underlying media is not redistributed, and users obtain it

Table S2: **Loss-weight sensitivity on *OmniRetriever-Bench* R@1.** *OmniRetriever-7B* under two loss-weight configurations ( $\lambda_T, \lambda_D, \lambda_A$ ). Both configurations improve over the pairwise-CL ablation (31.08) on AVG-all. Differences between the two settings are at most 2.27 R@1 in every direction and within 0.71 R@1 on AVG-all. We use the uniform (1, 1, 1) setting for the released model.

| Direction         | (1, 1, 1)    | (50, 10, 1) | $\Delta$ |
|-------------------|--------------|-------------|----------|
| text→video        | 48.89        | 47.41       | −1.48    |
| video→text        | 46.30        | 45.08       | −1.22    |
| text→audio        | 14.20        | 13.78       | −0.42    |
| audio→text        | 11.92        | 11.42       | −0.50    |
| video→audio       | 25.46        | 25.12       | −0.34    |
| audio→video       | 24.99        | 25.73       | +0.74    |
| text→audio+video  | 45.37        | 43.42       | −1.95    |
| audio+video→text  | 44.47        | 42.20       | −2.27    |
| audio→text+video  | 23.45        | 24.64       | +1.19    |
| text+video→audio  | 26.07        | 25.78       | −0.29    |
| video→audio+text  | 52.49        | 50.45       | −2.04    |
| audio+text→video  | 54.47        | 54.52       | +0.05    |
| <b>AVG single</b> | <b>28.63</b> | 28.09       | −0.54    |
| <b>AVG dual</b>   | <b>41.05</b> | 40.17       | −0.88    |
| <b>AVG all</b>    | <b>34.84</b> | 34.13       | −0.71    |

Table S3: **Caption-word statistics of the training corpus.**

|                  | mean  | median | p90 | p99 | max    | min |
|------------------|-------|--------|-----|-----|--------|-----|
| Caption (#words) | 32.04 | 30     | 49  | 76  | 11,938 | 1   |

|         |    | target → |      |      |      |      |      |
|---------|----|----------|------|------|------|------|------|
|         |    | T        | V    | A    | AV   | AT   | TV   |
| ↑ query | T  | —        | T→V  | T→A  | T→AV | —    | —    |
|         | V  | V→T      | —    | V→A  | —    | V→AT | —    |
|         | A  | A→T      | A→V  | —    | —    | —    | A→TV |
|         | AV | AV→T     | —    | —    | —    | —    | —    |
|         | AT | —        | AT→V | —    | —    | —    | —    |
|         | TV | —        | —    | TV→A | —    | —    | —    |

Figure S1: ***OmniRetriever-Bench* retrieval directions.** 12 valid query→target cells across a  $6 \times 6$  matrix: 6 single↔single (top-left  $3 \times 3$ ), 3 single→dual (top-right), 3 dual→single (bottom-left). Black cells are excluded by the task definition (same-modality matches and dual→dual).

from the source platform via the released identifiers. Sample-identifier disjointness with the training subset is enforced via SHA-256 identifier comparison.

## G Extended Compression Analysis

Table S4 reports the per-direction breakdown under three operating points: the 3584-d fp32 default, a 1024-d int8 setting (uniform dimension downsampling followed by symmetric per-vector min-max int8 quantization), and a 512-d binary setting (uniform downsampling followed by sign hashing). The same random indices are applied to query and gallery, and we report the mean over 5 random seeds. All compression is applied post-hoc on the released *OmniRetriever-7B* embeddings without retraining.

Under uniform-downsample 14× byte compression to int8/1024, AVG-all drops from 34.84 to 27.19; under 224× binary compression it drops to 9.14. The audio-related dual-modal directions retain more of their full-precision accuracy than the text–video directions at int8/1024. We leave compression-aware training (MRL+QAT) for future work.

**Front-dim truncation sanity check.** As an alternative to random downsampling, we also keep the first  $d$  dimensions of the embedding (the naive “head-of-vector” fallback used by some MRL deployments) and re-apply the same int8/binary



Figure S2: **OmniRetriever-Bench** sample cards. Each card pairs a mid-clip video keyframe and the recorded audio waveform with the caption used as  $T$ . The displayed samples are everyday user-generated content covering cooking, scenery, pets, narration over still imagery, ambient music, and casual dialogue, representative of the natural mix of short-form video in the pool.

Table S4: **Per-direction R@1 of OmniRetriever-7B under post-hoc uniform downsampling and quantization** (mean over 5 random-dimension seeds,  $\text{std} \leq 0.6$  everywhere). All compression is applied without retraining or MRL fine-tuning. **Storage:** fp32 = 14,336 B; int8/1024 = 1,024 B; bin/512 = 64 B.

| Direction         | fp32         | int8/1024 | bin/512 |
|-------------------|--------------|-----------|---------|
| text→video        | 48.89        | 34.87     | 4.12    |
| video→text        | 46.30        | 27.44     | 2.03    |
| text→audio        | 14.20        | 8.60      | 0.91    |
| audio→text        | 11.92        | 5.43      | 0.45    |
| video→audio       | 25.46        | 23.11     | 11.46   |
| audio→video       | 24.99        | 22.94     | 11.59   |
| text→audio+video  | 45.37        | 30.29     | 3.93    |
| audio+video→text  | 44.47        | 27.98     | 2.91    |
| audio→text+video  | 23.45        | 21.18     | 8.63    |
| text+video→audio  | 26.07        | 24.30     | 10.40   |
| video→audio+text  | 52.49        | 48.47     | 24.81   |
| audio+text→video  | 54.47        | 51.72     | 28.47   |
| <b>AVG single</b> | 28.63        | 20.40     | 5.09    |
| <b>AVG dual</b>   | 41.05        | 33.99     | 13.19   |
| <b>AVG all</b>    | <b>34.84</b> | 27.19     | 9.14    |

quantization. Differences vs. random downsampling are small (+0.62 R@1 at int8/1024, +0.21 at bin/512 on AVG-all, within 1 to  $2\sigma$  of the random-sample spread), so the two schemes are substantively equivalent. We adopt the random-sample formulation throughout since it does not assume any particular basis ordering.

## H OmniRetriever-Bench Test-Clip Duration Statistics

Table S5 shows duration quantiles measured on the *OmniRetriever-Bench* test pool ( $N = 3,782$ ). The quantiles justify our 32-frame / 0.5 s video sampling and 8 s audio crop at inference time.

## I External Cross-Modal Retrieval Benchmarks

This section complements main paper Sections 4.2 and 4.3 (Tables 2 and 3) with per- $k$  and

NDCG@10 details for all six benchmarks (Tables S6 and S7).

The remainder of this section reports per- $k$  and NDCG@10 details for completeness.

**Evaluation protocol.** All six benchmarks use the same  $N \rightarrow N$  diagonal-retrieval setup as *OmniRetriever-Bench*:  $N$  (multimodal\_item, text\_caption) pairs, the multimodal side being video+audio (when the original clip carries audio) for the four video tasks and audio for the two audio tasks. Recall@ $k$  counts the gold target landing in top- $k$ ; NDCG@10 reduces to  $1/\log_2(\text{rank}+1)$  if  $\text{rank} \leq 10$  else 0 under the single-relevant-document assumption. We use each benchmark’s official evaluation annotations; the evaluated pool size  $N$  is shown in each result table.

Table S5: **Duration quantiles (seconds) on the *OmniRetriever-Bench* test pool.** ( $N = 3,782$ ). Audio is the original recorded waveform.

| Stream | mean | p90  | p99   | max   |
|--------|------|------|-------|-------|
| Video  | 3.25 | 6.13 | 16.17 | 30.16 |
| Audio  | 3.13 | 5.99 | 16.02 | 30.00 |

Table S6: **Full external video–text benchmark results.** R@1 / R@5 / R@10 / NDCG@10.  $T \rightarrow AV$  sets the text caption as query against the video–audio item gallery;  $AV \rightarrow T$  the reverse. *OmniRetriever-7B* runs zero-shot without per-task fine-tuning.

| Benchmark ( $N$ )                         | T→AV  |       |       |         | AV→T  |       |       |         |
|-------------------------------------------|-------|-------|-------|---------|-------|-------|-------|---------|
|                                           | R@1   | R@5   | R@10  | NDCG@10 | R@1   | R@5   | R@10  | NDCG@10 |
| <i>OmniRetriever-7B (ours, zero-shot)</i> |       |       |       |         |       |       |       |         |
| MSR-VTT (1000)                            | 47.60 | 74.00 | 82.80 | 64.38   | 43.70 | 71.80 | 80.60 | 61.49   |
| MSVD (480)                                | 66.88 | 92.71 | 97.08 | 82.02   | 63.33 | 88.75 | 95.00 | 79.59   |
| DiDeMo (315)                              | 46.03 | 75.24 | 84.76 | 64.31   | 45.08 | 73.65 | 80.00 | 63.13   |
| VATEX (2494)                              | 57.98 | 89.05 | 94.63 | 76.80   | 54.97 | 87.21 | 93.18 | 74.68   |

## J Triple-Geometry Definition

For the analysis in Table 10, we use the *triple cosine score*  $\bar{c}_3(i) = \frac{1}{3}(\mathbf{z}_i^{(T)\top} \mathbf{z}_i^{(V)} + \mathbf{z}_i^{(T)\top} \mathbf{z}_i^{(A)} + \mathbf{z}_i^{(V)\top} \mathbf{z}_i^{(A)})$  on the *OmniRetriever-Bench* held-out pool of 3,782 triples and report both the *intra-sample* mean  $\mathbb{E}[\bar{c}_3^{\text{intra}}]$  (matched  $T, V, A$  from sample  $i$ ) and the *inter-sample* mean  $\mathbb{E}[\bar{c}_3^{\text{inter}}]$  (mismatched:  $T_i, V_{i+1}, A_{i+1}$ ).

## K Cross-Backbone Training Details

To test whether fusion-as-teacher distillation is specific to the WAVE-7B backbone, we re-train *OmniRetriever* on Omni-Embed-Nemotron-3B (Xu et al., 2025), an architecturally different unified embedder with an LLM trunk over Qwen2.5-Omni-Thinker, bidirectional attention, a mean-pool embedding, no all-layer fusion head, and  $\sim 3$  B parameters. The backbone is frozen except for LoRA adapters ( $r=16, \alpha=32$ ) on every LLM decoder layer’s q/k/v/o projections, matching the LoRA budget used on WAVE-7B. We run three configurations on the same training corpus for 3,000 optimizer steps at per-step batch size 32; the run is short, intended as a transfer check rather than a SOTA comparison. The configurations are PAIRWISE ( $\mathcal{L}_A$  only),  $\mathcal{L}_D + \mathcal{L}_A$ , and full *OmniRetriever* ( $\mathcal{L}_T + \mathcal{L}_D + \mathcal{L}_A$ ). Results under the main paper’s 12-direction protocol are reported in Table 8.

## L Caption-Paraphrase Robustness on *OmniRetriever-Bench*

*OmniRetriever-Bench* captions are human-corrected starting from a Gemini 3.0 Pro draft (Appendix F), so they are distinct from any caption in the training corpus. To verify that *OmniRetriever-7B*’s retrieval scores do not depend on specific caption surface forms, we paraphrase every *OmniRetriever-Bench* gold caption with a held-out LLM (gemini-2.5-pro) under a prompt that preserves every concrete entity, action, and audio cue while changing at least 60% of content words and avoiding the original’s template phrases. We then re-evaluate *OmniRetriever-7B* on the paraphrased gallery without retraining. Table S8 reports the two text-anchored bench directions ( $T \rightarrow AV$ ,  $AV \rightarrow T$ ), where caption surface form has the largest impact.

The drop is modest, consistent with the paraphrase merely lowering the caption-style alignment of the text encoder rather than degrading the underlying multimodal grounding. The audio-anchored directions ( $A \rightarrow T$ ,  $A \rightarrow TV$ ) on *OmniRetriever-Bench* only use the rewritten text as the retrieval *target* (not as the query modality); we observe a similar  $\sim 1$  R@1 drop on those directions when the rewritten captions are placed in the gallery.

Table S7: **Full external audio–text benchmark results.** R@1 / R@5 / R@10 / NDCG@10 for *OmniRetriever-7B* under zero-shot evaluation.

| Benchmark ( $N$ )                         | T→A          |              |              |              | A→T          |              |              |              |
|-------------------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                                           | R@1          | R@5          | R@10         | NDCG@10      | R@1          | R@5          | R@10         | NDCG@10      |
| <i>OmniRetriever-7B (ours, zero-shot)</i> |              |              |              |              |              |              |              |              |
| Clotho (1045)                             | <b>19.14</b> | <b>43.73</b> | <b>56.46</b> | <b>36.06</b> | <b>16.08</b> | <b>38.37</b> | <b>51.00</b> | <b>31.68</b> |
| SoundDescs (1000)                         | <b>25.00</b> | <b>52.70</b> | <b>66.40</b> | <b>43.95</b> | <b>20.70</b> | <b>46.50</b> | <b>58.80</b> | <b>37.98</b> |

Table S8: **Paraphrase-robustness on *OmniRetriever-Bench*.** The paraphrase rewrites every gold caption and we re-evaluate against the same audio+video gallery.  $A \leftrightarrow V$ .

| Direction                      | Original | Paraphrased | $\Delta$ |
|--------------------------------|----------|-------------|----------|
| <i>OmniRetriever-7B (ours)</i> |          |             |          |
| $T \rightarrow AV$             | 45.37    | 43.81       | −1.56    |
| $AV \rightarrow T$             | 44.47    | 43.39       | −1.08    |

## M $\mathcal{L}_T$ Seed-Stability Analysis

The aggregate +0.24 R@1 gain from adding  $\mathcal{L}_T$  on top of  $\mathcal{L}_D + \mathcal{L}_A$  (main paper Table 5) is close to the seed noise floor on aggregate, so we re-run both configurations under three seeds {42, 43, 44} and report the per-seed difference. Table S9 shows the result.

Table S9: **Per-seed  $\mathcal{L}_T$  ablation on *OmniRetriever-Bench* AVG-all R@1.** We report  $\mathcal{L}_D + \mathcal{L}_A$  and full *OmniRetriever* ( $\mathcal{L}_T + \mathcal{L}_D + \mathcal{L}_A$ ) under three seeds, plus the per-seed delta. The  $\mathcal{L}_T$  contribution is positive in every seed (+0.21 to +0.27), and the spread between seeds is small ( $\leq 0.08$  R@1 in either column), so the +0.24 aggregate gain is above seed noise.

| Seed | $\mathcal{L}_D + \mathcal{L}_A$ | full- <i>OmniRetriever</i> | $\Delta_{\mathcal{L}_T}$ |
|------|---------------------------------|----------------------------|--------------------------|
| 42   | 34.60                           | 34.84                      | +0.24                    |
| 43   | 34.56                           | 34.77                      | +0.21                    |
| 44   | 34.63                           | 34.90                      | +0.27                    |
| mean | 34.60                           | 34.84                      | <b>+0.24</b>             |
| std  | 0.04                            | 0.07                       | 0.03                     |

The seed std on either column is  $\leq 0.07$  R@1, much smaller than the +0.24 delta, so the  $\mathcal{L}_T$  gain on aggregate is well outside the noise envelope. The per-seed delta is positive in all three seeds; with only three seeds, we treat this as a consistency check rather than a formal significance test. The per-direction picture in Table 5 confirms the same story:  $\mathcal{L}_T$  adds +2 to +3 R@1 on the four hardest  $A \leftrightarrow V$  routes, well above per-direction noise. We conclude that  $\mathcal{L}_T$  is a real but small aggregate effect that primarily redistributes capacity toward