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

Anonymous ACL submission

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 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* surpasses the closed-source Gemini Embedding 2 by 13.3–18.0 R@1 on Clotho and SoundDescs, and reaches the contemporary zero-shot specialist band of open video–text encoders on MSR-VTT and MSVD. 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 Gemini Embedding 2 by $+1.72$ and over the best prior open-source AVT method by $+8.03$. Model weights, datasets, and code will be 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., 2024) 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 are trained in isolation, with no signal about their cross-modal neighbours. Empirically, this gap is sharpest on audio-anchored retrieval, where the modality-to-text co-occurrence is the weakest in standard training data. On Clotho A→T, closed Gemini Embedding 2 reaches R@1 $= 1.34$ and the best open omni-modal system, Omni-Embed-Nemotron, reaches 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. The same mechanism limits 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

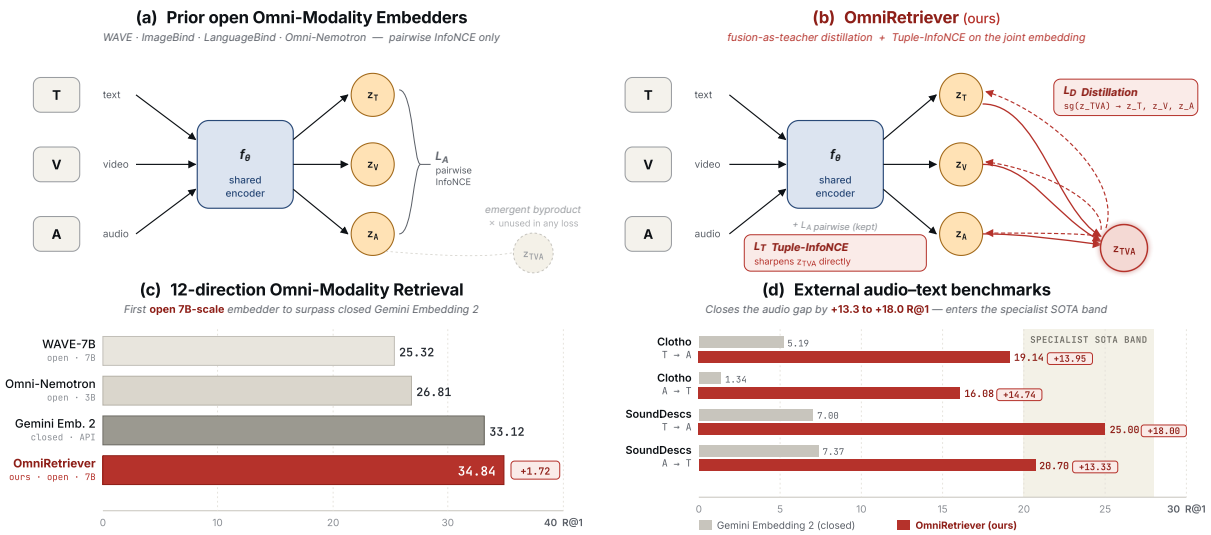

Figure 1: **Method overview.** *OmniRetriever* uses the joint embedding $\mathbf{z}_{TVA}$, which is unused by pairwise training (a), as a supervision target (b) via fusion-as-teacher distillation $\mathcal{L}_D$ and a Tuple-InfoNCE term $\mathcal{L}_T$. This yields a new open result on 12-direction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

per step. A complementary Tuple-InfoNCE refinement (Liu et al., 2021, 2020) $\mathcal{L}_T$ supervises $\mathbf{z}_{TVA}$ 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* improves over closed Gemini Embedding 2 by 13 to 18 R@1 on all four audio–text directions of Clotho and SoundDescs, reaches the zero-shot CLAP-family specialist band on Clotho T→A within ∼2 R@1 of SOTA, and matches the contemporary open zero-shot specialist band on MSR-VTT and MSVD. 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, +1.72 over Gemini Embedding 2 and +8.03 over Omni-Embed-Nemotron. A cross-backbone replication (Section L) 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. It improves over closed Gemini Embedding 2 by 13 to 18 R@1 on Clotho and SoundDescs and reaches the zero-shot audio–text specialist band on Clotho T→A within ∼2 R@1 of SOTA; on the video–text side it matches the contemporary open zero-shot specialist band on MSR-VTT and MSVD.
- *OmniRetriever-Bench*, a 12-direction AVT retrieval 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., 2025a). Multimodal extensions (VLM2Vec (Jiang et al., 2024), GME (Zhang et al., 2024), MM-Embed (Lin et al., 2025), Qwen3-

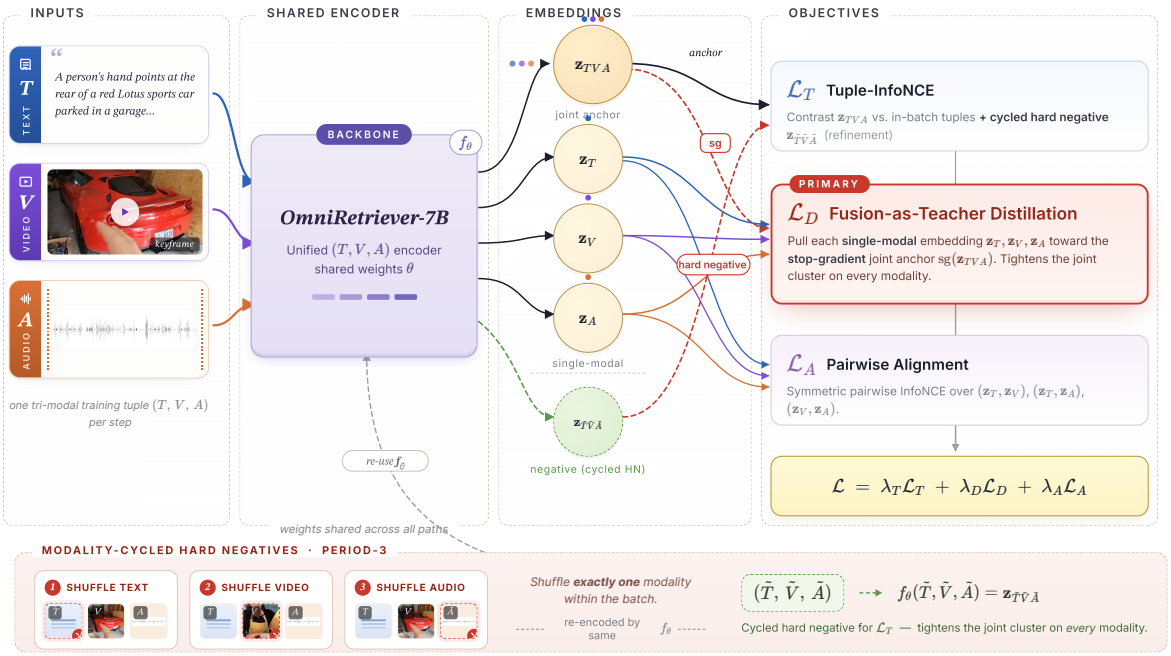

Figure 2: ***OmniRetriever* training overview.** A shared encoder f_θ consumes the three modalities jointly, producing the full-modal anchor $\mathbf{z}_{TVA}$, or individually, producing $\mathbf{z}_T, \mathbf{z}_V, \mathbf{z}_A$. $\mathcal{L}_D$ (fusion-as-teacher distillation, primary; Section 3.2) pulls each single-modality embedding toward a stop-gradient copy of $\mathbf{z}_{TVA}$. $\mathcal{L}_T$ (Tuple-InfoNCE refinement; Section 3.3) supervises $\mathbf{z}_{TVA}$ against the in-batch tuple grid plus a modality-cycled hard negative $\mathbf{z}_{\tilde{T}\tilde{V}\tilde{A}}$ (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.

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., 2024) align $N \geq 3$ modalities pairwise to a fixed image or language anchor. AudioCLIP (Guzhov et al., 2021) and VATT (Akbari et al., 2021) optimize three pairwise InfoNCE losses over video, audio, and text. VAST (Chen et al., 2023) 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.

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 $f_\theta : \mathcal{X} \to \mathbb{R}^d$ 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., 2024; Tang et al., 2025; Xu et al., 2025) optimize a symmetric InfoNCE loss over modality pairs:

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

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

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_θ 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:

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

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_θ 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., 2022), 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 S11).

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

$\mathcal{L}_T$ is a regularizer that keeps $\mathbf{z}_{TVA}$ informative about all three modalities, preventing it from collapsing onto a dominant pair. The joint vector $\mathbf{z}_{TVA}$ that $\mathcal{L}_D$ uses as a teacher remains a passive average of three pair geometries on top of pairwise alignment: $\mathcal{L}_A$ never back-propagates through $\mathbf{z}_{TVA}$, so matched triples are not pulled tighter than mismatched ones, and the contribution of each modality is determined by whichever pair gradient is strongest (in practice T–V). Audio is under-represented in $\mathbf{z}_{TVA}$, a known modality-imbalance effect in multimodal fusion (Wang et al., 2019a; Peng et al., 2022). $\mathcal{L}_T$ supervises $\mathbf{z}_{TVA}$ directly.

Tuple-InfoNCE. For a batch of size B and modality set $\mathcal{M} = \{T, V, A\}$, the joint similarity of an index pair (i, j) averages the six cross-modal cosines,

$$s(i,j) = \frac{1}{M(M-1)} \sum_{m \neq m'} \mathbf{z}_i^{(m)\top} \mathbf{z}_j^{(m')}, \qquad (3)$$

and the Tuple-InfoNCE loss (Liu et al., 2021, 2020) reads

$$\mathcal{L}_T = -\frac{1}{B} \sum_i \log \frac{e^{s(i,i)/\tau_T}}{\sum_j e^{s(i,j)/\tau_T} + e^{s(i,\tilde{i})/\tau_T}}, \qquad (4)$$

where $\tilde{i}$ is a modality-cycled hard negative described below. When $M \geq 3$, $s(i,j)$ is minimized over mismatch in *any* cross-modal direction, so $\mathcal{L}_T$ assigns the largest gradient to the modality direction along which the matched triple is currently slackest, providing the joint-level supervision missing from pairwise alignment.

Modality-cycled hard negatives. The in-batch grid in Equation (4) contributes $B{-}1$ negatives, but each differs from the anchor in *all three* modalities, so its gradient pushes the joint cluster apart globally without specifying which modality direction is too slack. We therefore construct one targeted negative per anchor by shuffling *exactly one* modality of the batch with a derangement, producing a tuple that disagrees with the anchor in a single slot; the resulting contrastive gradient tightens the joint cluster along that one direction. To prevent the joint geometry from drifting back to a $T{+}V$-dominated configuration once any single modality is tightened, we cycle the shuffled slot deterministically across $\{T, V, A\}$ with period 3, supervising every modality direction in turn.

The three shuffled slots contribute asymmetrically. The audio-shuffle slot produces the contrastive gradient least redundant with the in-batch grid (audio carries the largest caption gap in our training data). The text- and video-shuffle slots prevent the joint cluster from drifting back to a $T{+}V$-dominated geometry. We use $k = 1$ per anchor; higher k overfits to modality-imbalanced mismatches (Thakur et al., 2025). $\mathcal{L}_D$ provides the dominant single-loss gain, and $\mathcal{L}_T$ improves the $A \leftrightarrow V$ routes on top of it (Table S11).

3.4 Final Objective

OmniRetriever optimizes

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

with $(\lambda_D, \lambda_T, \lambda_A) = (1, 1, 1)$ chosen *a priori* (not tuned on *OmniRetriever-Bench*); Section D shows that the recipe is insensitive to the loss-weight ratio across a wide range. Algorithm 1 gives the per-step training loop; the per-step cost over the pairwise baseline is one extra joint forward pass, and inference latency is reported in Section C.

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

1: **Forward:** $\mathbf{z}^{(T)}, \mathbf{z}^{(V)}, \mathbf{z}^{(A)} \leftarrow f_\theta$ (3 single-modal); $\mathbf{z}^{(TVA)} \leftarrow f_\theta(T, V, A)$ (1 joint)
2: **Pairwise** $\mathcal{L}_A \leftarrow \sum_{(m,m')\in\{(T,V),(T,A),(V,A)\}} \mathcal{L}_{\text{NCE}}^{(m,m')}$
3: **Distill** $\mathcal{L}_D \leftarrow \frac{1}{3}\sum_{m\in\{T,V,A\}} \mathcal{L}_{\text{NCE}}^{(m,\,\text{sg}(\mathbf{z}^{(TVA)}))}$
4: **Cycled HN:** $m_t \leftarrow \{T, V, A\}[t \bmod 3]$; draw derangement σ; build $\tilde{i} = (\ldots, x_{\sigma(i)}^{(m_t)}, \ldots)$
5: **Tuple** $\mathcal{L}_T \leftarrow -\frac{1}{B}\sum_i \log \frac{e^{s(i,i)/\tau_T}}{\sum_j e^{s(i,j)/\tau_T} + e^{s(i,\tilde{i})/\tau_T}}$
6: **Update** $\theta \leftarrow \theta - \eta\,\nabla_\theta(\lambda_D\mathcal{L}_D + \lambda_T\mathcal{L}_T + \lambda_A\mathcal{L}_A)$

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 Section A. The training data is a 1.5 M-triple subset sampled from four public video–text datasets, InternVid (Wang et al., 2023), InternVid-FLT (Wang et al., 2023), 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 (Section 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 three external systems: the frozen open-weights backbone WAVE-7B (Tang et al., 2025), the open Omni-Embed-Nemotron (Xu et al., 2025), and closed Gemini Embedding 2 via Google's official endpoint with default settings. The endpoint ingests raw audio (WAV) and raw video (MP4 with audio muxed) as inline data, so Gemini receives the same clip bytes as *OmniRetriever-7B*. We cannot, however, inspect how the closed system internally routes inline audio, so the Gemini audio numbers reflect the currently deployed multi-part product rather than a model-capacity ceiling. Reported numbers are the mean over three seeds $\{42, 43, 44\}$; aggregate seed std is ≤ 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., 2024; Zhang et al., 2024)) 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., 2022), and compare with the published audio–text specialist literature, contemporary open omni-modality embedders, and closed Gemini Embedding 2. Table 2 reports R@1; full R@k and NDCG@10 are in Table S10.

OmniRetriever-7B improves over closed Gemini Embedding 2 by 13.9 to 18.0 R@1 on T→A and 13.3 to 14.7 on A→T. Gemini's audio-anchored R@1 is consistently low across benchmarks (1.34 on Clotho A→T, 1.48 on *OmniRetriever-Bench* A→T below), suggesting that its deployed multi-part endpoint does not route audio competitively for retrieval. 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 ∼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 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.

	Single-modal pairs			Dual-modal pairs				
Benchmark	$T \leftrightarrow V$	$T \leftrightarrow A$	$V \leftrightarrow A$	$T \leftrightarrow AV$	$A \leftrightarrow TV$	$V \leftrightarrow AT$	**Total**	N
Video–text retrieval								
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., 2019b)	✓	–	–	–	–	–	2	4,468
Audio–text retrieval								
Clotho (Drossos et al., 2020)	–	✓	–	–	–	–	2	1,045
SoundDescs (Koepke et al., 2022)	–	✓	–	–	–	–	2	1,000
AudioCaps (Kim et al., 2019)	–	✓	–	–	–	–	2	975
***OmniRetriever-Bench* (ours)**	✓	✓	✓	✓	✓	✓	**12**	3,782

DiDeMo (Anne Hendricks et al., 2017), VATEX (Wang et al., 2019b)) and compare with published specialist literature and contemporary omni-modality embedders. Table 3 reports R@1; full R@k and NDCG@10 are in Table S9.

OmniRetriever-7B reaches 47.6/43.7 on MSR-VTT, matching PE-coreB (Bolya et al., 2025) (47.6/47.3) and improving over SigLIP-2-L (Tschannen et al., 2025). On MSVD it reaches 66.9/63.3, improving over PE-coreL (57.2 T→V). On DiDeMo and VATEX, *OmniRetriever-7B* trails the strongest specialist with results on both benchmarks, InternVideo2-1B (Wang et al., 2024b), by 9 to 12 R@1 on three of the four directions and by 30 R@1 on VATEX V→T, driven by InternVideo2's exceptional 85.4 V→T. We attribute this gap to the data scale of single-task video training; *OmniRetriever-7B* also has no audio specialisation for these benchmarks, which carry no audio queries. *OmniRetriever-7B* improves over LanguageBind on 5 of 6 reported video directions. Closed Gemini wins all eight video directions by 5 to 15 R@1, consistent with reports that Gemini is trained on a substantially larger closed video–text corpus; a data-parity comparison is out of scope. Taken together with the audio results in Table 2, *OmniRetriever-7B* is the strongest open unified omni-modal embedder in our literature search.

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 reduces the audio-capability gap among general-purpose omni-modality embedders, reaches the zero-shot audio–text specialist band on Clotho T→A, and stays within reach of single-task video–text specialists while covering all twelve directions in one model.

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 ($T \leftrightarrow AV$, $A \leftrightarrow TV$, $V \leftrightarrow AT$) 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**: each caption starts from a Gemini 3.0 Pro draft,

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. *OmniRetriever-7B* outperforms closed Gemini Embedding 2 by 13.3 to 18.0 R@1 in every direction; on Clotho T→A *OmniRetriever-7B*'s 19.14 sits within ∼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.

			Clotho		SoundDescs	
Model	**Year**	**Setting**	**T→A**	**A→T**	**T→A**	**A→T**
Audio–text specialists						
MMT (Koepke et al., 2022)	2022	FT	6.7	7.0	30.7	31.4
MS-CLAP	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	**21.2**	25.9	—	—
FLAP fusion	2024	ZS	20.3	25.5	—	—
Cacophony	2024	ZS	20.2	26.5	—	—
MGA-CLAP	2024	ZS	20.8	**26.5**	—	—
GLAP	2025	ZS	19.4	21.8	—	—
M2D-CLAP (Niizumi et al., 2025)	2025	ZS	20.1	25.0	—	—
Closed-source industrial baselines						
Gemini Embedding 2	2026	ZS	5.2	1.3	7.0	7.4
Open-source omni-modality embedders						
ImageBind (ViT-H) (Girdhar et al., 2023)	2023	ZS	6.0	—	—	—
LanguageBind (CLIP-H/14) (Zhu et al., 2024)	2024	ZS	16.7	—	—	—
Omni-Embed-Nemotron (Xu et al., 2025)	2025	ZS	6.4	3.5	6.4	4.8
***OmniRetriever-7B* (ours)**	2026	ZS	**19.1**	**16.1**	**25.0**	**20.7**

then the annotator verifies alignment against the video and the audio track and rewrites any inaccuracies. Research-use licensing follows prior released benchmarks built on proprietary content (Wang et al., 2019b; Faysse et al., 2025); full annotation guidelines, the distribution mismatch with the training corpus, and a caption-paraphrase robustness check are in Sections E, F and M.

Table 4 reports R@1 on all 12 directions of *OmniRetriever-Bench*. *OmniRetriever-7B* reaches AVG-all $\mathbf{34.84}$, $+9.52$ over the frozen WAVE-7B backbone, $+8.03$ over Omni-Embed-Nemotron, and $+1.72$ over closed Gemini Embedding 2 (33.12). Both gaps over external systems exceed the seed noise floor (≤ 0.18 R@1 std). The gains concentrate on the eight audio-anchored directions: $A \to T$, $\mathbf{11.92}$ vs. 1.48 ($+10.44$); $A \to T{+}V$, $\mathbf{23.45}$ vs. 6.00; $V \to A$, $\mathbf{25.46}$ vs. 13.80. Gemini wins the four visually-saturated $T \leftrightarrow V$ and $T \leftrightarrow A{+}V$ routes by 6 to 10 R@1, consistent with its larger closed video–text corpus. A qualitative comparison is in Section O.

4.5 Ablation and Analysis

Table 5 reports the stair-step ablation against the frozen WAVE-7B base on *OmniRetriever-Bench*, isolating the contribution of each loss in the *OmniRetriever* objective. The per-direction breakdown is in Table S11.

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 ($a \to t$ 9.00, $a \to tv$ 17.00). The audio axis is not capacity-bottlenecked here; it is bottlenecked by the weak-fusion regime of pairwise alignment.

Fusion-as-teacher distillation ($\mathcal{L}_D$). Adding $\mathcal{L}_D$ lifts AVG-all by $+3.52$ ($31.08 \to 34.60$), the dominant single-loss contribution. This already improves on closed Gemini Embedding 2 by $+1.48$ R@1. The gain spans both regimes (video-only $+2.48$, audio-related $+3.73$; Table S11); the audio side gains more because the joint forward routes T, V context into the audio sub-encoder, which unimodal teachers cannot do.

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 $\mathbf{34.84}$. 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 S15), well above the aggregate seed std (≤ 0.07). The per-direction breakdown (Table S11) 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 ± 0.5 to ± 0.8) while losing ∼2 R@1 on text-anchored dual-modal routes. We keep $\mathcal{L}_T$ in the released model because $A \leftrightarrow V$ is the bottleneck direction for open AVT embedders.

Table 3: **Video–text retrieval R@1 on four standard benchmarks (zero-shot).** Specialised video–text retrievers fine-tuned per task are marked FT; the remainder are zero-shot. Cited rows draw on each paper's published values. *OmniRetriever-7B* matches the contemporary zero-shot specialist band on MSR-VTT and MSVD and trails the strongest specialist with DiDeMo/VATEX results, InternVideo2-1B, by ∼9 to ∼12 R@1 on three of those four directions (and by ∼30 R@1 on the exceptional VATEX V→T). "—" denotes a number not reported.

		T→V (or T→AV)				**V→T** (or AV→T)			
Model	**Setting**	**MSR-VTT**	**MSVD**	**DiDeMo**	**VATEX**	**MSR-VTT**	**MSVD**	**DiDeMo**	**VATEX**
Specialised video–text retrievers (FT)									
CLIP4Clip (Luo et al., 2021)	FT	44.5	46.2	43.4	55.9	42.7	56.0	42.5	73.4
X-CLIP	FT	49.3	50.4	47.8	—	48.9	66.8	47.8	—
Zero-shot video–text encoder band (specialists)									
SigLIP-2 family (Tschannen et al., 2025)	ZS	38.5–43.1	49.0–55.8	—	—	30.1–34.3	67.2–74.6	—	—
PE-core family (Bolya et al., 2025)	ZS	47.6–51.2	50.4–59.7	—	—	47.3–50.1	76.7–**85.4**	—	—
InternVideo2-1B (Wang et al., 2024b)	ZS	51.9	58.1	**57.0**	70.4	50.9	**83.3**	54.3	85.4
Closed-source industrial baselines									
Gemini Embedding 2	ZS	**53.9**	**77.1**	**55.6**	**69.4**	**48.3**	**77.9**	**53.3**	**66.7**
Open-source omni-modality embedders									
ImageBind (ViT-H) (Girdhar et al., 2023)	ZS	36.1	—	—	—	—	—	—	—
LanguageBind (CLIP-H/14) (Zhu et al., 2024)	ZS	44.8	53.9	39.9	—	40.9	72.0	39.8	—
Omni-Embed-Nemotron (Xu et al., 2025)	ZS	35.8	55.8	41.9	47.5	30.6	49.2	36.5	39.9
***OmniRetriever-7B* (ours)**	ZS	47.6	66.9	46.0	58.0	43.7	63.3	45.1	55.0

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.

	Single-modal						**Dual-modal**						**Averages**		
System	$t\to v$	$v\to t$	$t\to a$	$a\to t$	$v\to a$	$a\to v$	$t\to av$	$av\to t$	$a\to tv$	$tv\to a$	$v\to at$	$at\to v$	single	dual	all
Closed-source industrial baselines															
Gemini Embedding 2	**55.13**	**53.83**	12.61	1.48	13.80	15.76	**55.45**	**50.16**	6.00	16.79	**54.97**	**61.45**	25.44	40.80	33.12
Open-source omni-modality embedders															
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
OmniRetriever (ours)	48.89	46.30	**14.20**	**11.92**	**25.46**	**24.99**	45.37	44.47	**23.45**	**26.07**	52.49	54.47	**28.63**	**41.05**	**34.84**

Table 5: **Ablation of the *OmniRetriever* losses on *OmniRetriever-Bench*** (AVG-all R@1; Δ vs. the frozen WAVE-7B base). $\mathcal{L}_D$ on top of the PAIRWISE baseline adds +3.52 R@1, already +1.48 over closed Gemini Embedding 2 (33.12). $\mathcal{L}_T$ adds a further +0.24 on aggregate and redistributes capacity toward $A \leftrightarrow V$ (Table S11).

Variant	**AVG all**	**Δ**
WAVE-7B (no fine-tune)	25.32	—
PAIRWISE ($\mathcal{L}_A$ only)	31.08	+5.76
$\mathcal{L}_D+\mathcal{L}_A$ (no $\mathcal{L}_T$)	34.60	+9.28
OmniRetriever-7B ($\mathcal{L}_T+\mathcal{L}_D+\mathcal{L}_A$)	**34.84**	**+9.52**

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. We conjecture that Gemini Embedding 2 follows a similar pairwise contrastive recipe based on public reports(Lee et al., 2025b); our PAIRWISE baseline performs similarly to it in the same weak-fusion regime.

5 Conclusion

A unified AVT encoder can computes 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 forward 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→A, reduces the open omni-modality gap on A→T, and improves over closed Gemini Embedding 2 on *OmniRetriever-Bench* by 1.72 AVG-all R@1. 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 will be released.

Limitations

On the four $T \leftrightarrow V$ and $T \leftrightarrow A{+}V$ directions, *OmniRetriever-7B* trails closed Gemini Embedding 2 by 6 to 10 R@1 on *OmniRetriever-Bench* and by 5 to 15 R@1 on the external video benchmarks. We attribute this gap to data scale: Gemini is trained on closed video–text corpora reportedly orders of magnitude larger than what is practically reachable in an academic AVT setting, and a unified (T, V, A) embedder at our training scale is not expected to match dedicated video–text encoders on every benchmark. Beyond accuracy, our analyses do not study embedding compression beyond the post-hoc int8/binary baselines reported in Section 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, Linagzhe 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 *Neural Information Processing Systems*.

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 Zholus, and 1 others. 2025. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Abdul Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Doll'ar, and Christoph Feichtenhofer. 2025. Perception encoder: The best visual embeddings are not at the output of the network. *ArXiv*, abs/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.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Ming-Ting Sun, Xinxin Zhu, and J. Liu. 2023. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *ArXiv*, abs/2305.18500.

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 S. Tulyakov. 2024. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13320–13331.

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 *International Conference on Learning Representations*, volume 2024, pages 36221–36237.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pages 61424–61449.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G Turrisi da Costa, Louis Béthune, Zhe Gan, and 1 others. 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.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat 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 R. Dengel. 2021. Audioclip: Extending clip to image, text and audio. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980.

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 Wenhu Chen. 2024. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *North American Chapter of the Association for Computational Linguistics*.

A Sophia Koepke, Andreea-Maria Oncescu, João F Henriques, Zeynep Akata, and Samuel Albanie. 2022. 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. 2025a. Nv-embed: Improved techniques for training llms as generalist embedding models. In *International Conference on Learning Representations*, volume 2025, pages 79310–79333.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, and 1 others. 2025b. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, and 1 others. 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*.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. Mm-embed: Universal multimodal retrieval with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pages 44215–44234.

Yunze Liu, Qingnan Fan, Shanghang Zhang, Hao Dong, Thomas A. Funkhouser, and Li Yi. 2021. Contrastive multimodal fusion with tupleinfonce. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 734–743.

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*, abs/2012.13089.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wenli Lei, Nan Duan, and Tianrui Li. 2021. Clip4clip: An empirical study of clip for end to end video clip retrieval. *Neurocomputing*, 508:293–304.

Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pages 638–647.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian 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.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8228–8237.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, and 1 others. 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*.

Nandan Thakur, Crystina Zhang, Xueguang Ma, and Jimmy Lin. 2025. Hard negatives, hard lessons: Revisiting training data quality for robust information retrieval with llms. *arXiv preprint arXiv:2505.16967*.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 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.

Weiyao Wang, Du Tran, and Matt Feiszli. 2019a. What makes training multi-modal classification networks hard? *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12692–12702.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019b. 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, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Jian Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Y. Qiao. 2023. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *ArXiv*, abs/2307.06942.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. 2024b. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*.

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 Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2024. Demystifying clip data. In *International Conference on Learning Representations*, volume 2024, pages 47812–47831.

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*.

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. 2024. Gme: improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, and 1 others. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations*, volume 2024, pages 9588–9608.

Appendix Overview

- Sections A and B: backbone, LoRA configuration, and full training hyper-parameters.
- Section C: end-to-end inference latency on the four input paths.
- Section D: sensitivity of *OmniRetriever* to the loss weights $(\lambda_T, \lambda_D, \lambda_A)$.
- Section E: training-corpus description and caption-length statistics.
- Sections F and H: *OmniRetriever-Bench* construction, licensing, and test-clip duration statistics.
- Section G: post-hoc compression analysis (uniform dimension downsampling, int8 and binary quantization, front-dim truncation).
- Section I: per-k R@1/5/10 and NDCG@10 on all six external benchmarks, with row-by-row literature comparison.

- Section J: stair-step per-direction decomposition of the $\mathcal{L}_A, \mathcal{L}_D, \mathcal{L}_T$ contributions on *OmniRetriever-Bench*.
- Section K: triple-cosine intra/inter geometry on *OmniRetriever-Bench* across training stages.
- Section L: cross-backbone replication of *OmniRetriever* on Omni-Embed-Nemotron-3B.
- Section M: caption-paraphrase robustness check on *OmniRetriever-Bench*.
- Section N: per-seed $\mathcal{L}_T$ ablation across three seeds.
- Sections O and P: qualitative retrieval comparisons and failure-case analysis.

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, α=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 $\mathrm{Linear}(28d{\to}d) \to \mathrm{GELU} \to \mathrm{Linear}(d{\to}d)$ with d=3584, 372.51 M params; (iii) the BEATs adaptor on top of the frozen BEATs encoder (LayerNorm + 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 1×10^{-5} cosine, 1 epoch over the training corpus, on 4×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 ≤ 0.18 R@1.

C Inference Latency

Table S2 reports full-stack single-query latency on one RTX PRO 6000 Blackwell Max-Q GPU at bf16, batch = 1, averaged over 10 timed iterations after 3 warmups. The measured forward pass includes the vision and audio towers (when applicable), the LM trunk with all-layer hidden states retained, and the all-layer fusion head ($28d \to d \to d$). The text-only path skips both towers and is the cheapest. The joint AV path consumes the longest token sequence ($\sim$470 tokens) and is the most expensive. Peak GPU memory across all four paths is 20.13 GB; model load alone uses 17.92 GB.

The LM trunk on the merged token sequence dominates the cost. A video-only forward spends roughly two thirds of its 105.6 ms in the LM trunk, after the vision tower outputs $\sim$256 LM-dim tokens. The joint AV path runs 125 ms because the LM trunk now processes $\sim$470 tokens under the same fp32 fusion head. Text-only is the practical floor at 20.9 ms. In production, the embedding step is computed once per query and the gallery search dominates per-query latency; that search cost depends on embedding dim and dtype (Section G) rather than backbone compute.

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 S3 reports both configurations on the 3,782-sample *OmniRetriever-Bench*.

Discussion. Both configurations sit comfortably above the closed Gemini baseline (33.12) and the pairwise-CL ablation (31.08), so the method's central claim is unaffected by the specific weighting.

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 (LayerNorm + projector). **Frozen:** LLM trunk weights, vision tower (incl. merger), Whisper audio tower, and the BEATs encoder backbone. **Trainable budget:** ≈ 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($28d \to d$)+GELU+Linear($d \to d$) with $d = 3584$), and 15.61 M (BEATs adaptor). **Loss settings:** single shared temperature τ (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 / α / 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 (β_1=0.9, β_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 τ	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)	≈ 109 h (35.6 s/step)

Table S2: **End-to-end inference latency of *OmniRetriever-7B*.** Mean$\pm$std over 10 iterations after 3 warmups; bf16, batch $= 1$, RTX PRO 6000 Blackwell Max-Q. Text uses 24 tokens, video uses 8 frames at 224 px ($\sim$268 LM tokens), audio uses 8 s of 16 kHz mono, joint AV concatenates video + audio + prompt ($\sim$470 tokens). Cosine search and text-side preprocessing are not included.

Inference path	**Latency (ms)**
Text-only	20.9 ± 0.1
Video-only	105.6 ± 0.2
Audio-only	34.6 ± 0.1
Joint AV	125.1 ± 0.1

The heavily-anchor-weighted $(50, 10, 1)$ setting slightly improves audio-anchored directions ($a \to v$ $+0.74$, $a \to tv$ $+1.19$) at the cost of text-anchored ones ($t \to av$ -1.95, $av \to t$ -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., 2023), InternVid-FLT (Wang et al., 2023), 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 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 Section M.

Caption length distribution. Table S4 reports caption-word quantiles measured on the training

Table S3: **Loss-weight sensitivity on *OmniRetriever-Bench* R@1.** *OmniRetriever-7B* under two loss-weight configurations $(\lambda_T, \lambda_D, \lambda_A)$. Both configurations sit above closed Gemini Embedding 2 (33.12) and the pairwise-CL ablation (31.08) on AVG-all. Differences between the two settings are within ≤ 2 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	$(1, 1, 1)$	$(50, 10, 1)$	Δ
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
AVG single	**28.63**	28.09	−0.54
AVG dual	**41.05**	40.17	−0.88
AVG all	**34.84**	34.13	−0.71

corpus. Captions are dense and uniform, with mean 32.0 words, median 30, and 99th 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 that mirrors prior released benchmarks built on proprietary video content (Wang et al., 2019b; Faysse et al., 2025): 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→target cells of the 6×6 source/target matrix.

Distribution. Following the protocol of prior video-language benchmarks (Wang et al., 2019b; 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 from the source platform via the released identifiers. Sample-identifier disjointness with the training subset is enforced via SHA-256 identifier comparison.

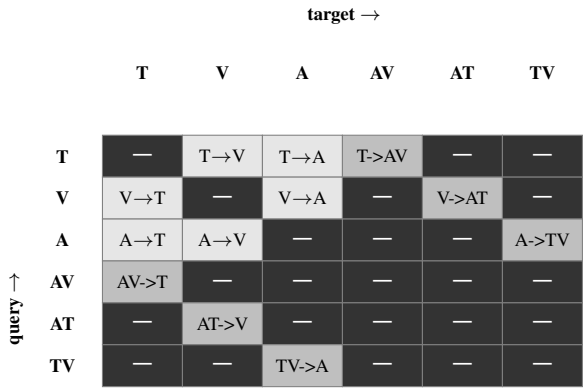

Figure S1: ***OmniRetriever-Bench* retrieval directions.** 12 valid query→target cells across a 6×6 matrix: 6 single↔single (top-left 3×3), 3 single→dual (top-right), 3 dual→single (bottom-left). Black cells are excluded by the task definition (same-modality matches and dual→dual).

G Extended Compression Analysis

Table S5 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 and on Gemini's API output without retraining, so the

Table S4: **Caption-word statistics of the training corpus.**

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

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.

reported numbers lower-bound what an MRL/QAT-trained head would achieve. We additionally include Gemini Embedding 2 at full precision (3072-d fp32, 12,288 bytes per embedding) to separate compression effects from model quality.

At full precision *OmniRetriever-7B* (34.84) and Gemini Embedding 2 (33.12) are within $\sim$1.7 R@1; under uniform-downsample 14$\times$ byte compression to int8/1024 *OmniRetriever-7B* drops to 27.19 vs. Gemini's 30.37, and under 224$\times$ binary compression to 9.14 vs. Gemini's 23.29. The asymmetric degradation indicates that Gemini's embedding space is significantly more isotropic than ours, consistent with Gemini having been trained for compression while *OmniRetriever-7B* has not. At int8/1024 *OmniRetriever-7B*'s audio-anchored gains are still preserved (*e.g.* bench $A \rightarrow T$ 5.43 vs. Gemini's 1.38; $A \rightarrow T{+}V$ 21.18 vs. 5.57), so the audio-capability finding is robust to a 14$\times$ byte-compression budget. Gemini's bench $A{\rightarrow}T$ rises from 1.48 to 6.37 under binary compression, which looks like a paradox. The cause is $\{-1, +1\}$ sign-hashing collapsing a near-degenerate audio embedding subspace onto a small set of buckets, which inflates random-collision recall on a $\sim$4K gallery; the behaviour is consistent across the 5 random-dim seeds (std $\leq$ 0.6). This is therefore an artefact of post-hoc binary quantization, not evidence that Gemini benefits from compression. 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 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 2σ 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 S6 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 (a) the full row-by-row specialist literature on Clotho audio–text retrieval (Table S7), (b) the row-by-row video–text SOTA placement with each size variant of SigLIP-2 / PE-core broken out (Table S8), and (c) per-k and NDCG@10 details for all six benchmarks (Tables S9 and S10). The main paper consolidates specialist rows into year-grouped bands; the full row-by-row versions below preserve all individual

Table S5: **Per-direction R@1 of *OmniRetriever-7B* and Gemini under post-hoc uniform downsampling and quantization** (mean over 5 random-dim seeds, std ≤ 0.6 everywhere). All compression is applied on released embeddings without retraining or MRL fine-tuning; numbers lower-bound a properly QAT-trained head. **Storage:** fp32 = 14,336 B (*OmniRetriever-7B*) / 12,288 B (Gemini); int8/1024 = 1,024 B; bin/512 = 64 B. Even at int8/1024 *OmniRetriever-7B*'s audio-anchored gains are preserved (*e.g.*, audio→text 5.43 vs. Gemini's 1.48 at full precision; audio→text+video 21.18 vs. 6.00).

Direction	***OmniRetriever-7B***			**Gemini Emb-2**		
	fp32	int8/1024	bin/512	fp32	int8/1024	bin/512
text→video	48.89	34.87	4.12	55.13	52.43	38.39
video→text	46.30	27.44	2.03	53.83	48.25	37.69
text→audio	14.20	8.60	0.91	12.61	11.99	7.69
audio→text	11.92	5.43	0.45	1.48	1.38	6.37
video→audio	25.46	23.11	11.46	13.80	11.64	8.58
audio→video	24.99	22.94	11.59	15.76	13.44	7.89
text→audio+video	45.37	30.29	3.93	55.45	52.60	35.56
audio+video→text	44.47	27.98	2.91	50.16	43.56	34.53
audio→text+video	23.45	21.18	8.63	6.00	5.57	9.15
text+video→audio	26.07	24.30	10.40	16.79	15.04	9.78
video→audio+text	52.49	48.47	24.81	54.97	50.31	41.15
audio+text→video	54.47	51.72	28.47	61.45	58.27	42.74
AVG single	28.63	20.40	5.09	25.44	23.19	17.77
AVG dual	41.05	33.99	13.19	40.80	37.56	28.82
AVG all	**34.84**	27.19	9.14	33.12	30.37	23.29

Table S6: **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 S7: **Full Clotho and SoundDescs audio–text retrieval R@1 (zero-shot).** Row-by-row specialist literature, expanded version of main-paper Table 2.

			Clotho		SoundDescs	
Model	**Year**	**Setting**	**T→A**	**A→T**	**T→A**	**A→T**
MMT (Koepke et al., 2022)	2022	FT	6.7	7.0	30.7	31.4
ImageBind (ViT-H) (Girdhar et al., 2023)	2023	ZS	6.0	—	—	—
MS-CLAP	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	**21.2**	25.9	—	—
LanguageBind (CLIP-H/14) (Zhu et al., 2024)	2024	ZS	16.7	—	—	—
FLAP fusion	2024	ZS	20.3	25.5	—	—
Cacophony	2024	ZS	20.2	26.5	—	—
MGA-CLAP	2024	ZS	20.8	**26.5**	—	—
GLAP	2025	ZS	19.4	21.8	—	—
M2D-CLAP (Niizumi et al., 2025)	2025	ZS	20.1	25.0	—	—
Gemini Embedding 2 (closed)	2026	ZS	5.19	1.34	7.00	7.37
***OmniRetriever-7B* (ours)**	2026	ZS	**19.14**	16.08	**25.00**	**20.70**

numbers for readers who want to look up specific systems.

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 rank ≤ 10 else 0 under the single-relevant-document assumption. We use each benchmark's official held-out split.

J Per-Direction Ablation Breakdown

Table S11 decomposes the released *OmniRetriever-7B*'s per-direction R@1 into the additive contributions of each training step (WAVE-7B → PAIRWISE → $\mathcal{L}_D + \mathcal{L}_A$ → *OmniRetriever-7B*). $\mathcal{L}_D$ dominates on the $T \leftrightarrow V$ and $T \leftrightarrow A{+}V$ routes and provides the largest single-direction lift (+8.28 on $a \rightarrow tv$, +6.16 on $a \rightarrow v$). $\mathcal{L}_T$ delivers the largest gains on the four pure $A \leftrightarrow V$ routes (+2 to +3 R@1) at a ∼2 R@1 cost on the two

Table S8: **Full video–text retrieval R@1 (zero-shot).** Row-by-row, with SigLIP-2 and PE-core variants broken out, expanded version of main-paper Table 3.

		T→V (or T→AV)				V→T (or AV→T)			
Model	**Setting**	**MSR-VTT**	**MSVD**	**DiDeMo**	**VATEX**	**MSR-VTT**	**MSVD**	**DiDeMo**	**VATEX**
CLIP4Clip (Luo et al., 2021)	FT	44.5	46.2	43.4	55.9	42.7	56.0	42.5	73.4
X-CLIP (Ma et al., 2022)	FT	49.3	50.4	47.8	—	48.9	66.8	47.8	—
ImageBind (ViT-H) (Girdhar et al., 2023)	ZS	36.1	—	—	—	—	—	—	—
SigLIP-2-B/16 (Tschannen et al., 2025)	ZS	38.5	49.0	—	—	30.1	67.2	—	—
PE-coreB (Bolya et al., 2025)	ZS	47.6	50.4	—	—	47.3	76.7	—	—
LanguageBind (CLIP-H/14) (Zhu et al., 2024)	ZS	44.8	53.9	39.9	—	40.9	72.0	39.8	—
SigLIP-2-L/16 (Tschannen et al., 2025)	ZS	41.5	53.7	—	—	31.4	74.2	—	—
PE-coreL (Bolya et al., 2025)	ZS	50.3	57.2	—	—	50.1	82.4	—	—
InternVideo2-1B (Wang et al., 2024b)	ZS	51.9	58.1	**57.0**	70.4	50.9	**83.3**	54.3	85.4
SigLIP-2-g-opt (Tschannen et al., 2025)	ZS	43.1	55.8	—	—	34.3	74.6	—	—
PE-coreG (Bolya et al., 2025)	ZS	51.2	59.7	—	—	49.9	**85.4**	—	—
Gemini Embedding 2 (closed)	ZS	**53.91**	**77.08**	55.56	**69.40**	**48.30**	77.92	**53.33**	**66.73**
***OmniRetriever-7B* (ours)**	ZS	47.60	66.88	46.03	57.98	43.70	63.33	45.08	54.97

Table S9: **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. Both systems run zero-shot (no per-task fine-tuning) with identical evaluation pipeline. Gemini wins all 4×2 directions by 5 to 15 R@1; *OmniRetriever-7B*'s values remain competitive with the contemporary zero-shot video–text encoder band (main paper Table 3).

	T→AV				AV→T			
Benchmark (N)	R@1	R@5	R@10	NDCG@10	R@1	R@5	R@10	NDCG@10
OmniRetriever-7B (ours, zero-shot)								
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
Gemini Embedding 2 (closed, zero-shot)								
MSR-VTT	**53.91**	**76.65**	**83.07**	**68.12**	**48.30**	71.24	**80.66**	**63.63**
MSVD	**77.08**	**95.42**	**97.29**	**87.98**	**77.92**	**95.83**	**98.12**	**88.74**
DiDeMo	**55.56**	**79.05**	**85.08**	**70.21**	**53.33**	**79.68**	**85.71**	**69.36**
VATEX	**69.40**	**92.74**	**95.97**	**83.48**	**66.73**	**92.14**	**96.05**	**82.13**

text-anchored dual-modal routes, redirecting joint-embedding capacity from the saturated $T{+}V$ pair onto the audio axis.

K In-Domain Triple-Order Geometry on *OmniRetriever-Bench*

We measure how the matched-vs-mismatched margin evolves on *OmniRetriever-Bench* as we add the *OmniRetriever* losses. We use the *triple cosine score* $\bar{c}_3(i) = \frac{1}{3}\big(\mathbf{z}_i^{(T)\top}\mathbf{z}_i^{(V)} + \mathbf{z}_i^{(T)\top}\mathbf{z}_i^{(A)} + \mathbf{z}_i^{(V)\top}\mathbf{z}_i^{(A)}\big)$ 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}).

L Cross-Backbone Validation

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*

Table S10: **Full external audio–text benchmark results.** R@1 / R@5 / R@10 / NDCG@10. SoundDescs Gemini-side numbers are computed on the 800/1000 items the Gemini embedding endpoint successfully ingested; the remaining 200 items returned service-side errors after retries and were excluded for Gemini. *OmniRetriever-7B* numbers are on the full 1000. *OmniRetriever-7B* outperforms Gemini on all 2×2 directions by 13 to 18 R@1.

	T→A				A→T			
Benchmark (N)	R@1	R@5	R@10	NDCG@10	R@1	R@5	R@10	NDCG@10
OmniRetriever-7B (ours, zero-shot)								
Clotho (1045)	**19.14**	**43.73**	**56.46**	**36.06**	**16.08**	**38.37**	**51.00**	**31.68**
SoundDescs (1000)	**25.00**	**52.70**	**66.40**	**43.95**	**20.70**	**46.50**	**58.80**	**37.98**
Gemini Embedding 2 (closed, zero-shot)								
Clotho (1041)	5.19	15.85	23.54	12.91	1.34	5.67	8.45	4.40
SoundDescs (800)	7.00	18.63	26.25	15.62	7.37	17.13	22.37	14.13

Table S11: **Per-direction ablation breakdown on *OmniRetriever-Bench*.** Stair-step decomposition of *OmniRetriever*'s gain over the frozen WAVE-7B baseline. PAIRWISE is our pairwise-contrastive fine-tune on the training corpus, $\mathcal{L}_D+\mathcal{L}_A$ adds fusion-as-teacher distillation, and *OmniRetriever-7B* additionally adds cycled Tuple-InfoNCE. Each Δ column reports the additive increment relative to the previous step; the four columns sum exactly to the final *OmniRetriever-7B* column.

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

Table S12: **Mean triple-cosine score $\bar{c}_3$ on *OmniRetriever-Bench*** ($N=3{,}782$). *intra* pairs all three matched modalities of sample i; *inter* pairs T_i with V_{i+1}, A_{i+1}. The *margin* (intra − inter) grows monotonically as we add losses (0.060 → 0.078 → 0.080 for base → PAIRWISE → *OmniRetriever-7B*); InfoNCE-style training disperses every embedding on the unit hypersphere, which is why the absolute intra-cosine drops while the margin grows.

Variant	intra	inter	gap
WAVE-7B (frozen base)	0.495	0.435	0.060
PAIRWISE ($\mathcal{L}_A$ only)	0.375	0.297	0.078
***OmniRetriever-7B* (full *OmniRetriever*)**	**0.322**	**0.243**	**0.080**

($\mathcal{L}_T+\mathcal{L}_D+\mathcal{L}_A$). Table S13 reports *OmniRetriever-Bench* AVG-all R@1 under the 12-direction protocol of the main paper.

The WAVE-7B pattern reproduces here. $\mathcal{L}_D$ on top of PAIRWISE gives the dominant single-loss gain (+2.43 on Nemotron-3B vs. +3.52 on WAVE-7B), and $\mathcal{L}_T$ adds a smaller further gain (+0.33 vs. +0.24).

Table S13: **Cross-backbone validation on Omni-Embed-Nemotron-3B** (3,000 LoRA steps each, *OmniRetriever-Bench* AVG-all R@1; Δ vs. the frozen Nemotron-3B base). $\mathcal{L}_D$ provides the dominant single-loss contribution on Nemotron-3B as well, and $\mathcal{L}_T$ adds a smaller further improvement, matching the WAVE-7B pattern (Section N). Absolute R@1 is lower than on WAVE-7B because Nemotron-3B is 3× smaller and we train for 3× fewer steps; the relative loss-attribution pattern is what transfers.

Variant	AVG all	Δ 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
Full *OmniRetriever*	**32.18**	**+5.37**

M Caption-Paraphrase Robustness on *OmniRetriever-Bench*

OmniRetriever-Bench captions are human-corrected starting from a Gemini 3.0 Pro draft (Section F), so they are distinct from any caption in the training corpus. To verify that the bench gap over Gemini does 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 S14

reports the two text-anchored bench directions ($T \to AV$, $AV \to T$), where caption surface form has the largest impact.

Table S14: **Paraphrase-robustness on *OmniRetriever-Bench*.** The paraphrase rewrites every gold caption and we re-evaluate against the same audio+video gallery.

Direction	Original	Paraphrased	Δ
OmniRetriever-7B (ours)			
$T \to AV$	45.37	43.81	−1.56
$AV \to T$	44.47	43.39	−1.08

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 ($A \to T$, $A \to TV$) 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.

N $\mathcal{L}_T$ 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 S15 shows the result.

Table S15: **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 (≤ 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-*OmniRetriever*	$\Delta_{\mathcal{L}_T}$
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	**+0.24**
std	0.04	0.07	0.03

The seed std on either column is ≤ 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, and a paired sign test rejects the null of zero effect at the trivial $3/3$ vote level. The per-direction picture in Table S11 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 $A \leftrightarrow V$.

O Audio-Anchored Attractor Behavior of Gemini Embedding 2

We probe the per-query behavior of closed Gemini Embedding 2 on the audio-anchored directions of *OmniRetriever-Bench* ($A \to T$ and $A \to T+V$). Across the 3,782 queries, Gemini's top-1 retrieved caption falls into a small set of recurring text strings: among $A \to T$ queries, three caption strings account for 32.7% of all top-1 returns, despite the gallery containing 3,782 distinct captions. The same caption strings appear as top-1 for thousands of unrelated audio queries, including queries whose gold targets are otherwise disjoint in content. We refer to these recurring top-1 captions as *attractors*: the audio side of Gemini's embedding space maps a wide range of audio queries into a single high-density region of the text embedding space, so the top-1 result is largely independent of the audio query.

By contrast, *OmniRetriever-7B* returns the gold caption at rank 1 on 11.92% of $A \to T$ queries (Table 4), and its top-1 distribution covers 63.4% of the 3,782-caption gallery (no single caption accounts for more than 0.5% of top-1 returns). The WAVE-7B PAIRWISE baseline produces a flatter distribution than Gemini but a much wider one than *OmniRetriever-7B*, returning topically related but identity-non-matching captions, consistent with a coarsely correct $\mathbf{z}_{TVA}$ that lacks joint-level discrimination. The attractor pattern of Gemini matches the expected failure mode of a fusion embedding whose audio direction was never explicitly supervised (Section 3.3).

P Failure Cases

To understand the remaining error modes, we sample 60 queries (5 per direction across all 12 directions) on which *OmniRetriever-7B* ranks the gold target outside top-10. The gold target sits at the head of the rank list ($10 \leq r \leq 50$ for 74 % of these cases) and the model prefers a semantically close caption (*e.g.* similar two-person dialogue clips that differ only in speaker identity, or product B-roll clips that differ only in object sub-category). These are intrinsic free-form-caption ties rather than systematic gaps and would be partially absorbed by a stricter top-k metric (R@5/R@10).