Omni-Embed-Mini: Binding Modalities Without Forgetting via Dense Distillation

Anonymous ACL submission

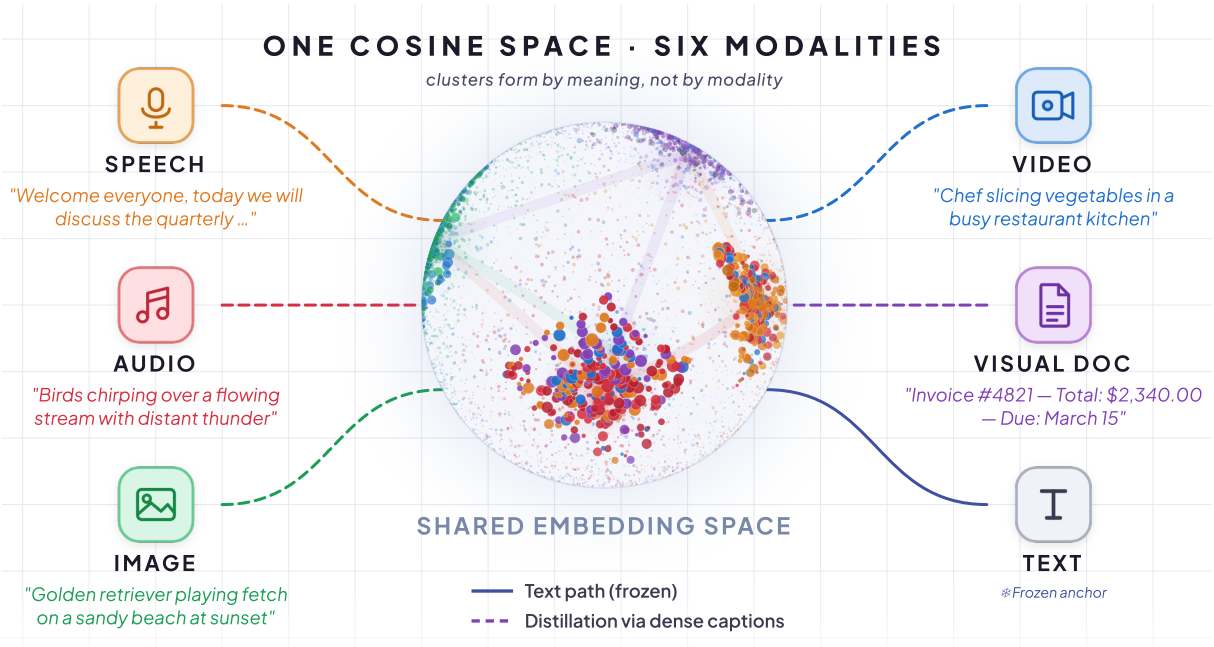


Figure 1: **Omni-Embed-Mini overview.** Omni-Embed-Mini maps speech, audio, image, video, visual-document, and text inputs into a single cosine embedding space, where semantically related examples cluster together regardless of modality. The model keeps the text encoder as a frozen anchor and aligns non-text modalities to this shared space through distillation from dense captions.

Abstract

Extending a text embedding model to new modalities typically *degrades* text retrieval quality, and existing omni-modal embedders compensate with multi-billion parameters. We present **Omni-Embed-Mini**, a 0.9B-parameter model that maps text, speech, audio, images, video, and visually-rich documents into a single shared cosine space *without* updating any text-side parameter. Our key insight is that the teacher signal requires no separate model: each media sample is paired with a dense cascaded caption, and the teacher target is simply the frozen backbone's own embedding of that caption. Because teacher and student share the same backbone weights, they inhabit byte-identical geometry, and lightweight projectors plus phased LoRA adapters on the modality encoders suffice for alignment. Training combines a Matryoshka SigLIP contrastive loss with an online hybrid hard-negative miner whose negatives sharpen as the encoder improves. The recipe scales unchanged to a 2.3B variant by swapping in a native vision-language backbone. Omni-Embed-Mini-0.9B *preserves* the backbone's text retrieval quality (49.50 nDCG@10 on MTEB-v2 BEIR-8, matching our reproduction of the standalone backbone) while extending it to five additional modalities, and is ∼2.7–9.5× smaller than every competing omni embedder.

1 Introduction

A central goal of recent representation learning is a unified embedding space that binds diverse modalities (Girdhar et al., 2023; Zhu et al., 2024).

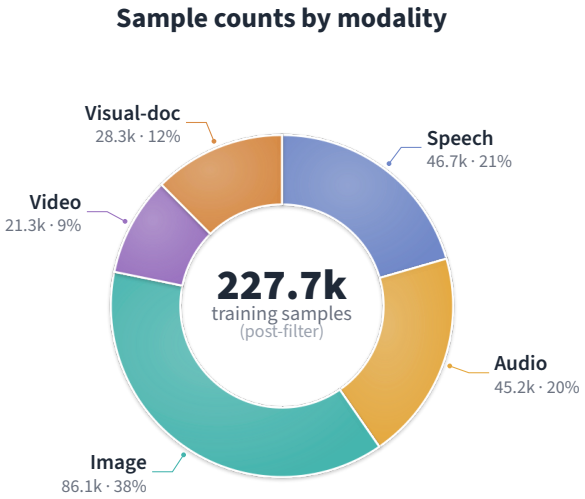


Figure 2: **Training data composition.** Sample-count distribution across the five non-text modalities in the caption-only slice of our omni-modal corpus, covering 227k effective rows after caption and duration filtering.

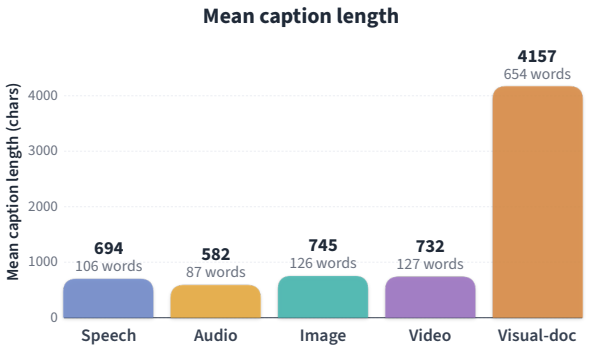


Figure 3: **Cascaded-caption length by modality.** Mean caption length per modality in the caption-only slice of our omni-modal corpus. Visual documents are $\sim 6\times$ longer because the expert-analyst captioner enumerates layout, headings, tables, and charts.

In practice, every attempt to jointly contrastive-train text with vision and audio exhibits two failure modes that have, so far, been treated as unavoidable. (*i*) *Catastrophic forgetting on the text side*: fine-tuning a strong text embedder against media inputs shifts the text geometry and degrades MTEB performance. Text is still the most critical modality for real-world search, making this degradation highly undesirable. (*ii*) *Modality bloat*: existing omni embedders such as LCO-Embedding-3B/7B (Xiao et al., 2025), BidirLM-Omni-2.5B (Boizard et al., 2026), and NVIDIA's omni-embed-nemotron-3b (Xu et al., 2025) have scaled to 2.5B–8.9B parameters in search of enough capacity to absorb every modality, rendering them unsuitable for edge-deployed dynamic retrieval at scale. We term these two coupled failure modes the *multimodal expansion dilemma*: every additional modality results in an unavoidable trade-off: mitigating text-side catastrophic forgetting currently requires massive parameter bloat.

We argue that text should not be a co-learner; it should be an *immovable anchor*. BLIP-2 (Li et al., 2023) bridged two frozen towers with a small Q-Former, and LLaVA (Liu et al., 2023) showed that a lightweight projector suffices to align vision and language during alignment pretraining with a frozen LLM. We translate this principle to retrieval, drawing on the same logic Reimers and Gurevych (2020) used to align a new language to a frozen English sentence space: each media sample's teacher target is the EOS-pooled embedding of a dense cascaded caption, computed by the *identical frozen backbone* that processes the student. Student and teacher therefore inhabit byte-identical geometry; a modest projector and phased LoRA on the encoders close the modality gap. We call the resulting recipe **Omni-Embed-Mini**.

As a result, Omni-Embed-Mini establishes a new Pareto frontier for size-vs-performance. Despite being 2.7–9.5$\times$ smaller than current omni-modal embedders, our 0.9B model retains 100% of its foundational text retrieval accuracy while delivering competitive omni-modal retrieval.

Our contributions are as follows. (*i*) **Zero-regression omni embedding:** a 0.9B model (Table 1) preserves a 0.6B text backbone's MTEB-v2 score (49.50 nDCG@10, matching our standalone reproduction) while adding five modalities; the text inference path is byte-identical to the stock backbone. (*ii*) **Cascaded-caption self-distillation through a shared backbone:** teacher and student share weights, so cosine alignment in the native pooled hidden space suffices: no projection head, no auxiliary teacher, fully cacheable targets. (*iii*) **Matryoshka SigLIP + online hybrid hard-negative mining + phased LoRA:** a parameter-efficient recipe that scales unchanged to a 2.3B variant by swapping the backbone for a native vision-language model. (*iv*) To our knowledge, this is the **smallest open omni embedder** at submission time, covering text, speech, audio, image, video, and visually-rich documents in one cosine space.

2 Related Work

Per-modality embedders. *Text.* The MTEB benchmark family (Muennighoff et al., 2023;

Component	Params
Frozen text backbone (Qwen3-Embedding-0.6B)	595.78 M
Vision encoder (ViT from Qwen3.5-0.8B)	100.59 M
Whisper-small encoder	88.15 M
Dasheng-base encoder	85.46 M
Projectors and pooler	65.37 M
Total (Omni-Embed-Mini-0.9B)	**935.35 M**

Table 1: **Parameter breakdown** for Omni-Embed-Mini-0.9B. The 2.3B variant replaces the backbone and custom vision encoder with native Qwen3-VL-Embedding-2B ($\sim$2.10 B), and uses larger Whisper-large-v3 and Dasheng-0.6B audio encoders.

Thakur et al., 2021) has strong text encoders such as Qwen3-Embedding (Zhang et al., 2025), voyage-3 (Voyage AI, 2024), Conan-embedding-v2 (TencentBAC, 2025), GritLM (Muennighoff et al., 2024), inf-retriever (Yang et al., 2025), and LENS (Lei et al., 2025); we treat these as backbone as the text-side reference against which our zero-regression claim is measured. *Vision–language.* VLM2Vec-V2 (Meng et al., 2025), Qwen3-VL-Embedding (Li et al., 2026), RzenEmbed (Jian et al., 2025), Ops-MM-embedding (Alibaba Cloud OpenSearch-AI team, 2025), and seed-1.6 (ByteDance Seed, 2025) extend a VL backbone with a contrastive head but do not handle audio. *Audio.* CLAP (Wu et al., 2023b) and MS-CLAP (Elizalde et al., 2023) introduced contrastive audio-text alignment; Whisper (Radford et al., 2023) (weakly supervised) and Dasheng (Dinkel et al., 2024) (self-supervised) are strong audio encoders without an aligned text tower.

Omni-modal embedders. LCO-Embedding-3B/7B (Xiao et al., 2025), BidirLM-Omni-2.5B (Boizard et al., 2026) (no video), and NVIDIA's omni-embed-nemotron-3b (Xu et al., 2025) target a single space across text, image, video, and audio, but are $\sim$ 2.7-9.5$\times$ larger than Omni-Embed-Mini-0.9B and trade per-modality accuracy for coverage. Crucially, attempting to absorb all modalities via joint contrastive updating typically degrades foundational text-retrieval performance, a critical flaw for real-world deployment. (e.g., omni-embed-nemotron achieves only 16.87 Avg nDCG@10 on ViDoRe-V3; see Fig 5).

Anchored vs. contrastive joint alignment. ImageBind (Girdhar et al., 2023) uses *images* as a binding modality: a pretrained CLIP ViT-H image encoder serves as anchor, and per-modality encoders are trained with symmetric InfoNCE to match the image embedding of naturally co-occurring pairs (image-audio from video, image-depth from RGB-D, etc.). Cross-modal alignment emerges because each modality is independently pulled toward the shared image space. LanguageBind (Zhu et al., 2024) adopts the same anchored-alignment structure but substitutes *language* for images as the binding modality, freezing a text encoder and training per-modality encoders against it. AudioCLIP (Guzhov et al., 2022) takes a different joint-training approach, extending CLIP with a three-way contrastive loss across image, text, and audio. Omni-Embed-Mini shares LanguageBind's choice of text as anchor but departs from both ImageBind and LanguageBind in three ways: (*i*) alignment is *asymmetric distillation* (the frozen backbone embeds a dense caption as the teacher target) rather than symmetric contrast, so no encoder space drifts except the student's; (*ii*) pairing data is *manufactured* via dense captioning rather than requiring natural co-occurrence, removing ImageBind's reliance on modality-specific paired corpora and LanguageBind's dependence on curated video-language datasets; (*iii*) the text encoder is not frozen but is the *identical backbone* that serves both teacher and student, guaranteeing byte-identical text geometry and zero regression on text retrieval. Conceptually, this is the cross-modal analogue of (Reimers and Gurevych, 2020), who aligned a new language to a frozen English sentence space via distillation against translated sentences; we substitute "language" with "modality" and "translation" with "dense captioning," inheriting the corresponding zero-regression property on text.

Our recipe builds on several established techniques. ANCE (Xiong et al., 2020) establishes that periodically refreshing the hard-negative pool against the current encoder is essential for strong dense retrievers; our online hybrid miner is the multimodal-Matryoshka generalisation, with two FAISS-CPU (Johnson et al., 2019) indices on independent cadences. Matryoshka representation learning (Kusupati et al., 2022; OpenAI, 2024) enables a single trained embedding to be truncated post-hoc; we inherit MRL from the backbone and propagate it to all modalities through the distillation loss. SigLIP (Zhai et al., 2023) provides the pair-wise sigmoid contrastive loss that decouples positives from negatives at small batch size, and LoRA (Hu et al., 2022) provides parameter-efficient adapters on the modality encoders.

3 Method

3.1 Asymmetric modality alignment

Two design principles and one training rule govern Omni-Embed-Mini. **(P1) Backbone invariance.** The text backbone is never updated: no full fine-tuning, no LoRA on the backbone. The text path is byte-identical to the stock backbone at inference, preserving its calibrated text-retrieval quality by construction. **(P2) Cascaded-caption self-distillation.** Each media sample is paired with a dense caption; the embedding of that caption, obtained via the *same* frozen backbone, is the teacher target. Teacher and student share weights, so a cosine alignment loss in the native pooled hidden space suffices. **(T1) Alignment-first, then adaptation.** Phase 1 trains projectors only; Phase 2 (after 20% of steps) injects encoder LoRA. Activating LoRA earlier wastes encoder budget on noisy upstream projections.

3.2 Architecture

Backbone and pooling. A decoder-only transformer with causal attention; an EOS-pooling head extracts the last non-padding hidden state and L2-normalises it (no projection head). The 0.9B variant uses Qwen3-Embedding-0.6B (Zhang et al., 2025) (hidden 1024) with a ViT extracted from Qwen3.5-0.8B; the 2.3B variant uses Qwen3-VL-Embedding-2B (Li et al., 2026) (hidden 2048) with native vision.

Audio path: dual encoder, temporally interleaved. Whisper (Radford et al., 2023) (mel) and Dasheng (Dinkel et al., 2024) (raw waveform) run in parallel; each emits 128 projected tokens per 30 s chunk via a 1-D temporal down-projector with SwiGLU residual. The two streams are *interleaved in time*, $(w_1, d_1, w_2, d_2, \dots)$, yielding 256 tokens per chunk. Rather than a learned cross-attention bottleneck, temporal interleaving provides a structural inductive bias that encourages the backbone's causal attention to attend jointly to semantic speech content and acoustic texture within local context windows, at zero extra parameter cost.

Vision path. *Custom (0.9B):* ViT $\rightarrow$ 2×2 spatial merger $\rightarrow$ dimension projector $\rightarrow$ backbone hidden size; for video, a spatial pooler applies cross-attention against a learnable 196-query budget so video sequence length is independent of frame count. *Native (2.3B):* the backbone's own visual module handles image and video; audio is injected via a forward hook on the embedding layer. Image and video tokens carry multimodal RoPE; audio tokens are spliced as a contiguous temporal span and reuse text RoPE. Media embeddings are spliced into modality-tagged placeholder positions via a one-hot matmul so gradients flow through the projectors and (Phase 2) encoder LoRA.

3.3 Cascaded-caption self-distillation

Each training sample i is paired with a dense caption c_i generated by Qwen3-Omni-30B-A3B-Instruct, with the source dataset's ground-truth caption pinned in the system prompt as a grounding prior (§B). The teacher embedding t_i is the L2-normalised EOS-pooled output of the frozen backbone applied to c_i; because the backbone is frozen, t_i is constant across training and is computed once and cached on disk under a stable sample id. The student embedding s_i is the same EOS-pool applied to the token sequence with media spliced in. Because s_i and t_i share the backbone and pooling, they live in identical geometry; no auxiliary projection head is required.

3.4 Hybrid Matryoshka Contrastive Learning

Let $D = \{128, 256, 512, 1024\,(, 2048)\}$ denote the set of Matryoshka dimensions. For each $d \in D$, we truncate the embeddings to their first d coordinates and re-normalise them.

We compute a SigLIP pair-wise sigmoid loss (Zhai et al., 2023) over in-batch and hard-negative pairs, using a learned temperature τ and bias b:

$$\mathcal{L}_{\text{siglip}}^{(d)} = \mathcal{L}_{\text{siglip}}\big(\hat{s}^{(d)}, \hat{t}^{(d)}; \tau, b\big)$$

The Matryoshka contrastive loss aggregates the per-dimension losses with weights proportional to $1/\sqrt{d}$, thereby favouring aggressive truncations:

$$w_d = \frac{d^{-1/2}}{\sum_{d' \in D} d'^{-1/2}}, \qquad \mathcal{L}_{\text{contr}} = \sum_{d \in D} w_d \, \mathcal{L}_{\text{siglip}}^{(d)}$$

We further use cosine self-distillation to align the student and teacher representations across Matryoshka dimensions:

$$\mathcal{L}_{\text{dist}} = \sum_{d \in D} w_d \, \mathbb{E}_i \left[1 - \hat{s}_i^{(d)} \cdot \hat{t}_i^{(d)}\right]$$

The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{contr}} + \lambda \mathcal{L}_{\text{dist}}, \qquad \lambda = 1$$

Full equations and implementation details are provided in §G.

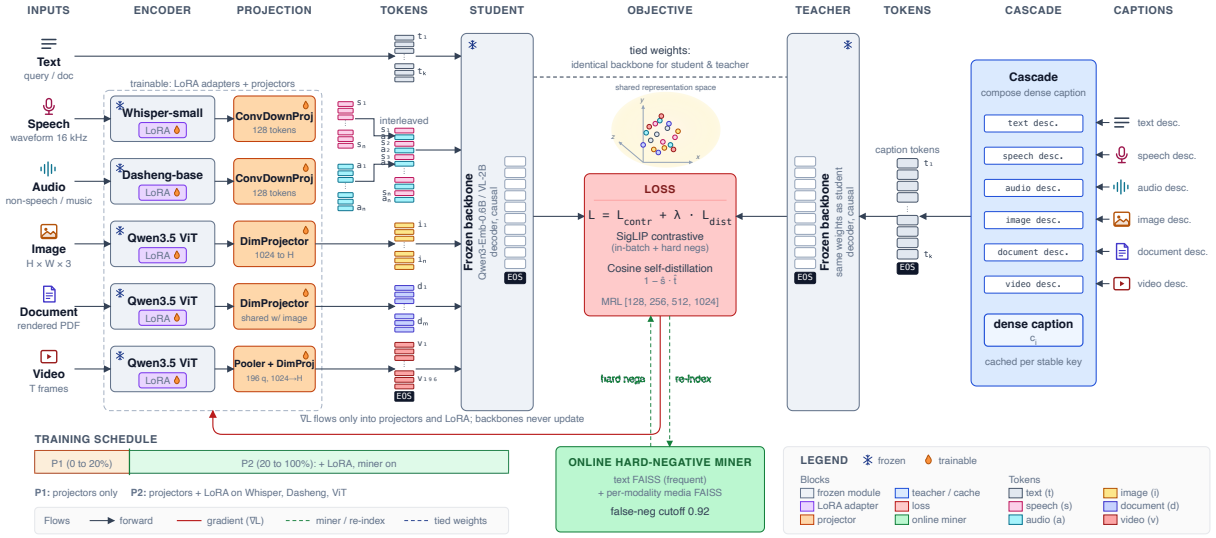


Figure 4: **Omni-Embed-Mini architecture.** Six student inputs (text, speech, audio, image, document, video) flow through a single *frozen* causal text backbone; a cascaded-caption teacher mirrors them through the same backbone. Plain text passes straight through unchanged, guaranteeing zero text-side regression. Speech is encoded by Whisper (mel) and audio by Dasheng (raw 16 kHz); each emits 128 projected tokens, *temporally interleaved* into a 256-token column. Image, document, and video share a Qwen3.5 ViT; video additionally passes a spatial pooler that collapses to a fixed 196-token budget. The EOS-pooled hidden state is the embedding (no projection head); the teacher path embeds the dense caption through the same backbone, with t_i cached to disk under a stable composite key.

3.5 Online hybrid hard-negative mining

Hard negatives are supplied by a background subsystem maintaining two FAISS-CPU (Johnson et al., 2019) inner-product indices, in the spirit of ANCE (Xiong et al., 2020). A lightweight *text* cycle ($\sim$10$\times$/epoch) re-embeds captions via the text-only path and retains the top-k neighbours per query. A more compute-intensive *media* cycle ($\sim$5$\times$/epoch) re-embeds samples per modality through the full model into per-modality indices. The miner shares the live model with the training loop, so each cycle returns negatives from the *current* representation: as the encoder improves, mined negatives become harder. Crucially, because the text branch never receives gradient, the text-side index is perfectly stable; only the media-side index drifts, and only toward harder negatives. This stability prevents *modality collapse* without ever destabilising text geometry. Concurrency, cache semantics, and failure handling are detailed in §D.

4 Training Data: Dense Caption Grounding

A central claim of this paper is that the success of frozen-backbone distillation depends critically on the *density* of the captioning signal. We therefore treat data preparation as a first-class methodological contribution rather than a logistical detail.

Source corpus. We train on the caption slice of our omni-modal caption corpus. From 590,858 total rows, filtering to caption-category rows across the speech, audio, image, video, and visual-doc configurations yields 242,080 single-turn samples; a subsequent duration filter (removing audio/video clips outside the supported length range) reduces this to 227,704 effective training rows (Fig. 2). The omni split (cross-modal multi-turn chat) and chat samples are excluded; their narrow answers are unsuitable distillation targets.

Cascaded caption pipeline. Every row is the output of a single low-temperature (T=0.7) call to Qwen3-Omni-30B-A3B-Instruct, with the source dataset's ground-truth caption interpolated as a "Reference caption" inside a modality-specific system prompt. This *grounds* the cascaded caption against the source rather than letting the captioner hallucinate. Visual-doc rows use an "expert document analyst" prompt with no reference caption (none exists). Full prompts and worked examples are in Appendix §B.

Dense captions are the *single most important data-side ingredient* of the recipe (as we demonstrate in §6.2): they expose more facets of each sample (objects, attributes, layout, temporal events, acoustic environment, OCR text) than short crowd-sourced captions, which act as a bag-of-words sig-

nal that encourages geometric collapse. We ablate this in Table 2, and worked qualitative examples per modality are in Appendix Table 9.

5 Experiments

5.1 Benchmarks

Style	Sp.	Aud.	Img.	Vid.	Overall
Original Captions	43.45	32.25	17.66	5.13	24.62
Dense Captions	**45.15**	**35.64**	**24.31**	**17.52**	**30.66**
Δ	+1.70	+3.39	+6.65	+12.39	+6.04

Table 2: **Dense vs. original captions** on the 0.9B configuration. *Original* denotes source-shipped captions, and *Dense* denotes our dense cascaded captions. Sp., Aud., Img., Vid., and Overall denote speech, audio, image, video, and the mean of the four, respectively. Speech/audio are MAEB task-mean splits; image/video are MMEB-V2 AvgImg/AvgVid.

We evaluate on four community benchmarks; full schemas in Appendix §A. **MTEB-v2** (Muennighoff et al., 2023; Thakur et al., 2021): text retrieval, the 8-task BEIR English subset (ArguAna, CQA-E/P/Pr, FiQA, NFCorpus, SCIDOCS, SciFact); metric nDCG@10. Role: *text-side regression test*. **MAEB** (El Assadi et al., 2026): 22 English audio tasks (retrieval, classification, clustering, multi-label, pair, reranking, zero-shot); mixed metrics. Role: cross-modal audio-to-text alignment plus audio coverage. **MMEB-V2** (Meng et al., 2025): 16 English tasks (10 image, 6 video); metric hit@1. **ViDoRe-V3** (Loison et al., 2026): 7 English visually-rich domains; metric nDCG@10.

English-only protocol. MMEB, MAEB, and ViDoRe are taken almost in full (16/18, 22/22, 7/7 supported English tasks); MTEB is reduced to BEIR-8 because the full v2 catalog has 1000+ task-language pairs. Each benchmark is re-validated by re-running its published anchor in our harness; per-benchmark validation deltas and known model quirks are in §A.

5.2 Setup

Hardware. 8× AMD Instinct MI210 (64 GB HBM2e each), bf16, DDP. ROCm-specific patches (audio I/O substitution, FAISS forced to CPU, BatchNorm fallback for CLAP) are summarised in Appendix §G.

Hyperparameters and cost. AdamW (wd 0.01, clip 1.0); 2% linear warm-up → cosine; LR 2×10^{-4} for audio projectors, 1×10^{-4} for vision projectors and encoder LoRA (Phase 2 only); LoRA r=16, α=32; phase switch at 20%; MRL dims $\{128, 256, 512, 1024\}$ (+2048 for 2.3B); $\lambda_{\text{dist}} = 1$. The 0.9B checkpoint is trained for 4,320 steps and the 2.3B for 3,231 steps. Full table in §G. For video evaluation on MMEB-V2, all omni-style models embed 8 frames as images and mean-pool, keeping comparisons uniform but capping achievable video performance (§H).

5.3 Main results

Fig. 5 summarises the headline result restricted to the omni class. At **0.9B**, our model is the smallest omni embedder by ∼2.7× relative to the next entry (BidirLM-Omni-2.5B) and ∼9.5× relative to LCO-Omni-7B; it leads the entire omni block on *Text* (49.50, the only omni embedder that preserves a strong text backbone exactly) and stays within a few points of much larger omni baselines on *Speech* and *Audio*. The **2.3B** variant, built on a frozen Qwen3-VL-Embedding-2B, reaches an overall-modality average of 47.55, leading the omni block on *Video* and *Visual-Doc* and placing second on *Image* at less than half the parameter count of the next-largest omni baseline. The all-class version of these panels (text-only, VL, audio/CLAP, and omni baselines together) is in Appendix Fig. 8; per-task tables are in Appendix §A.

The 0.9B variant scores lower on *Image* (24.31) and *Video* (17.52) because its vision encoder is extracted from a 0.8B backbone with a comparatively small ViT, trained on only ∼86K image and video rows. The 2.3B variant, which uses the native Qwen3-VL vision encoder, closes this gap substantially (54.09 image, 48.62 video), confirming that the bottleneck is vision-encoder capacity rather than the alignment recipe itself.

A notable side-effect of routing every modality through a shared text anchor is that within-modality and cross-modal retrieval emerges without any dedicated training signal. For instance, on the image-to-image tasks NIGHTS and FashionIQ (Tab 6), our 2.3B variant scores 68.00 and 34.50, leading the omni block on both despite never seeing an image-to-image objective. Similarly, audio and speech clustering and pair-classification tasks in MAEB (Tab 4, 5) operate entirely within-modality, yet our model performs competitively. We discuss this emergent alignment further in Appendix §E.

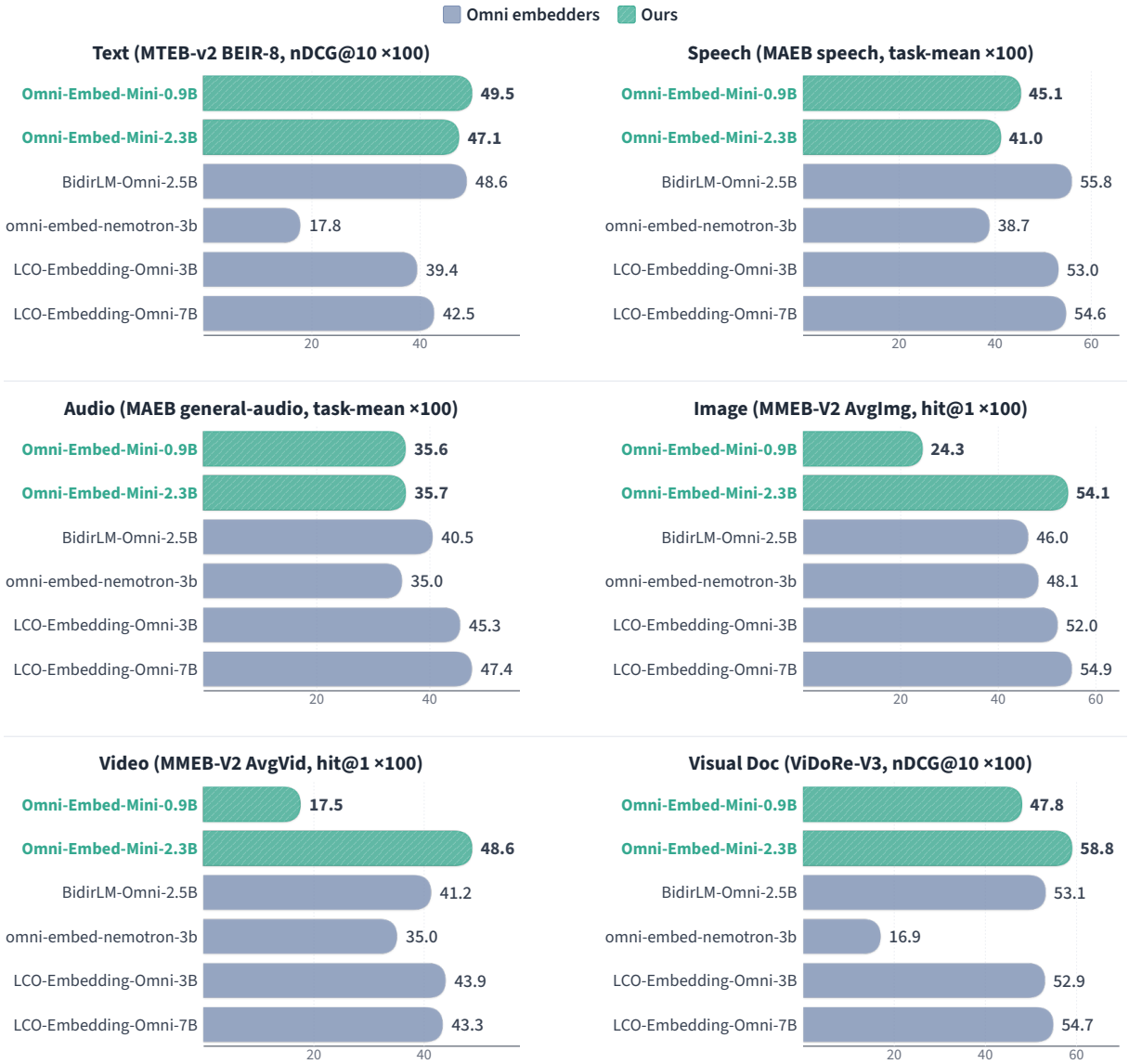


Figure 5: **Per-modality rankings, omni-only.** Same panels as Appendix Fig. 8 restricted to omni embedders. Both Omni-Embed-Mini variants beat existing omni baselines on text and speech at ∼2.7–9.5× fewer parameters; the 2.3B variant additionally inherits Qwen3-VL-Embedding-2B image/video/visual-doc strengths.

6 Ablations and Analysis

We isolate four design choices; frozen backbone, dense captions, hybrid mining, and Matryoshka propagation; through targeted ablations.

6.1 Zero-shot parity

On MTEB-v2 BEIR-8, Omni-Embed-Mini-0.9B scores **49.50**, identical to our standalone reproduction of the backbone (Qwen3-Embedding-0.6B). This is not a tuning artefact: the text-only inference path touches only the unchanged backbone weights, never the trained projectors or LoRA adapters. The text branch is byte-identical to the pretrained model. Per-task numbers are in Appendix Table 3.

6.2 Dense vs. original captions

Table 2 shows the smallest gap is on speech (+1.7), where source transcripts already carry most of the content; the largest is on video (+12.39), where short clip captions miss nearly every temporal event. The same effect motivated the dense recaptioning pipeline of Li et al. (2024) for VLM pre-training; we show it is equally decisive for retrieval.

6.3 Hybrid mining

Fig. 6 shows that, on most modalities and in overall average, the ordering is *text-only* < *no-mining* < *media-only* < *both*. Text-only mining alone slightly

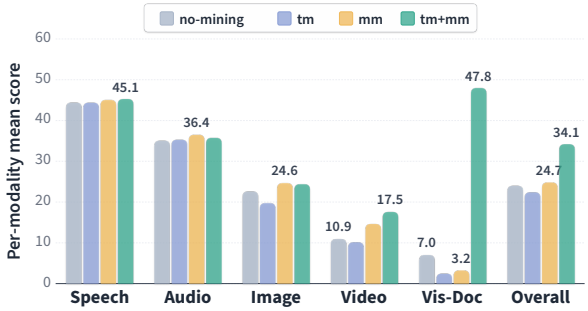


Figure 6: **Mining ablation.** Per-modality means on the 0.9B recipe across no-mining (in-batch only), text-only, media-only, and combined mining; text is invariant by construction.

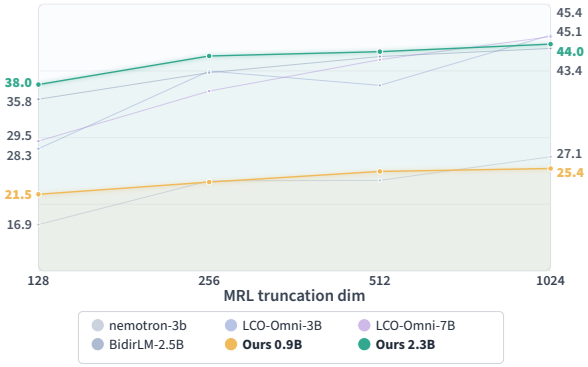


Figure 7: **MRL truncation behaviour.** AVG over per-modality probes (two retrieval tasks each from MTEB-v2, MAEB, MMEB-V2, ViDoRe-V3) as a function of truncation dimension $d \in \{128, 256, 512, 1024\}$. Both *Ours* curves are near-monotone and stay close to full-dimension performance even at the most aggressive 128-dim slice, while non-MRL omni baselines lose 35–38% over the same range.

hurts on several modalities (the student sees harder text negatives but no harder media ones, biasing the projector); media-only is generally a clean improvement; combining both yields the best overall score, though per-modality rankings occasionally swap between media-only and combined (e.g. audio and image). The visual-doc panel is especially decisive: every other configuration leaves the model below 7 nDCG@10 on ViDoRe-V3, while combined mining jumps to 47.8, a $\sim$7$\times$ swing attributable to joint hard-negative pressure. The two miners contribute complementary pressure: text-mining sharpens the caption-text geometry the teacher signal targets, while media-mining sharpens the modality-specific student geometry; dropping either leaves a slack axis along which the projector can drift. We use combined mining for the main recipe.

6.4 Matryoshka behaviour

Because MRL slicing acts on the backbone's pooled hidden state with no learned head, the Matryoshka behaviour at every dimension is inherited from the backbone (Kusupati et al., 2022; OpenAI, 2024). As shown in Fig. 7, the averaged-probe AVG drops by only $\sim$3.9 points (1024 $\rightarrow$ 128) for the 0.9B model and $\sim$6.1 for the 2.3B variant, so the most aggressive 8$\times$ slice keeps $\sim$85% of full-dimension quality. A 128-dim truncation reduces on-disk footprint 8$\times$: a 10M-item omni index fits in $\sim$5 GB rather than $\sim$40 GB, without sacrificing meaningful accuracy on any modality. Additional analyses (a SciFact text-vs-speech NDCG@10 convergence probe, emergent media-to-media alignment via the shared text axis, and the single-stage data-loader / phased optimiser schedule) are reported in §E.

7 Conclusion

We introduced **Omni-Embed-Mini**, a 0.9B omni-modal retrieval model that binds text, speech, audio, images, video, and visually-rich documents into a unified space by recasting cross-modal alignment as self-distillation through a shared, *frozen* text backbone. By keeping the text encoder fixed and aligning every other modality to its representation, we sidestep the catastrophic forgetting that typically plagues multi-modal fine-tuning and turn omni-modal training into a lightweight projector-plus-LoRA exercise. The model preserves its backbone's text retrieval quality while outperforming omni embedders $\sim$2.7$\times$ to 9.5$\times$ larger, and scales unchanged to a 2.3B variant competitive with baselines 2$\times$ to 4$\times$ its size. Because the backbone is frozen, future text-embedder improvements can be adopted by re-running only the projector and LoRA training, making the recipe a practical path for the community to ride the curve of rapidly improving text retrievers without retraining from scratch. Open threads include learned mixing of the dual audio-encoder tokens, a Mixture-of-Encoders variant that routes inputs to specialist branches, and replacing per-frame mean-pooling for video (§H) with a native spatial pooler that better preserves temporal and spatial structure. Beyond architecture, the most pressing community gap is evaluation: existing benchmarks each probe a single modality slice, and none evaluate cross-modal retrieval in arbitrary query$\rightarrow$corpus combinations across all six modalities simultaneously.

Limitations

Omni-Embed-Mini is a research prototype for omni-modal retrieval and is not validated for high-stakes deployment. Although the model provides efficient alignment across text, speech, audio, image, video, and visually-rich documents, it may still retrieve incorrect, biased, private, or misleading items when used in downstream systems. Therefore, it should not be used as a standalone decision system in sensitive domains without task-specific validation and human oversight.

The model's shared embedding space is constrained by the frozen text backbone used to preserve text-side retrieval performance. As a result, modality-specific phenomena that are difficult to verbalize, such as complex acoustic textures, abstract visual patterns, or fine-grained temporal cues, may not be fully represented. The cascaded-caption self-distillation pipeline also depends on teacher-generated captions, which may contain hallucinations, omissions, or biases that are inherited by the student model.

The audio pathway currently uses a fixed temporal interleaving of speech and environmental-audio tokens, which may be suboptimal for inputs dominated by only one audio type. In addition, the compact 0.9B and 2.3B model scales are optimized for efficiency, but may limit performance on complex reasoning-heavy multimodal retrieval tasks.

Finally, the evaluation is limited by the coverage of available benchmarks. Existing datasets do not yet provide a unified any-to-any retrieval benchmark across all six supported modalities, so our results may not generalize to all query–corpus combinations, languages, domains, data qualities, or real-world retrieval workflows.

Ethical Considerations

This work is intended for research and educational use in omni-modal representation learning and retrieval. Omni-Embed-Mini should not be used for autonomous decision-making in high-stakes domains such as medicine, law, finance, hiring, education, security, or public-sector services without appropriate validation and human supervision.

The main ethical risks include inaccurate retrieval, bias amplification, privacy leakage, and misuse. Since the model aligns multiple modalities into a shared embedding space, errors or biases from training data, teacher-generated captions, and pre-trained encoders may affect retrieval quality across languages, accents, visual appearances, environments, document types, and cultural contexts. Downstream users should conduct safety, fairness, privacy, robustness, and human-in-the-loop evaluations before applying the model in sensitive settings.

We also disclose our use of generative AI tools. Writing-support models were used for grammar, style, and clarity improvements. All technical claims, experimental results, analyses, and conclusions were authored, checked, and verified by the authors; any coding assistance was limited to boilerplate scaffolding, minor refactoring, or formatting support.

References

Alibaba Cloud OpenSearch-AI team. 2025. Ops-MM-embedding-v1-7B. Hugging Face model card.

Nicolas Boizard, Théo Deschamps-Berger, Hippolyte Gisserot-Boukhlef, Céline Hudelot, and Pierre Colombo. 2026. Bidirlm: From text to omnimodal bidirectional encoders by adapting and composing causal llms. *arXiv preprint arXiv:2604.02045*.

ByteDance Seed. 2025. seed1.6-embedding / seed-1.6-embedding-1215. https://seed1-6-embedding.github.io.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, and 16 others. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.

Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang, and Bin Wang. 2024. Scaling up masked audio encoder learning for general audio classification. *arXiv preprint arXiv:2406.06992*.

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.

Adnan El Assadi, Isaac Chung, Chenghao Xiao, Roman Solomatin, Animesh Jha, Rahul Chand, Silky Singh, Kaitlyn Wang, Ali Sartaz Khan, Marc Moussa Nasser, Sufen Fong, Pengfei He, Alan Xiao, Ayush Sunil Munot, Aditya Shrivastava, Artem Gazizov, Niklas Muennighoff, and Kenneth Enevoldsen. 2026. Maeb: Massive audio embedding benchmark. *arXiv preprint arXiv:2602.16008*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning

audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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.

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

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. Tedlium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *International conference on speech and computer*, pages 198–208. Springer.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Weijian Jian, Yajun Zhang, Dawei Liang, Chunyu Xie, Yixiao He, Dawei Leng, and Yuhui Yin. 2025. Rzenembed: Towards comprehensive multimodal retrieval. *arXiv preprint arXiv:2510.27350*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE transactions on big data*, 7(3):535–547.

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.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

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

Yibin Lei, Tao Shen, Yu Cao, and Andrew Yates. 2025. Enhancing lexicon-based text embeddings with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18986–19001.

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

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 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*.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, Yuyin Zhou, and Cihang Xie. 2024. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel Moreira, Bo Liu, Manuel Faysse, Céline Hudelot, and Gautier Viaud. 2026. Vidore v3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios. *arXiv preprint arXiv:2601.08620*.

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.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. 2025. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

OpenAI. 2024. New embedding models and API updates (text-embedding-3 with matryoshka representation). https://openai.com/blog/new-embedding-models-and-api-updates.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 4512–4525.

SB. 2025. gemini-flash-2.0-speech.

Min-Han Shih, Ho-Lam Chung, Yu-Chi Pai, Ming-Hao Hsu, Guan-Ting Lin, Shang-Wen Li, and Hung-yi Lee. 2023. Gsqa: An end-to-end model for generative spoken question answering. *arXiv preprint arXiv:2312.09781*.

TencentBAC. 2025. Conan-embedding-v2. https://huggingface.co/TencentBAC/Conan-embedding-v2.

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

Voyage AI. 2024. voyage-3-m-exp. https://huggingface.co/voyageai/voyage-3-m-exp.

Yijing Wu, SaiKrishna Rallabandi, Ravisutha Srinivasamurthy, Parag Pravin Dakle, Alolika Gon, and Preethi Raghavan. 2023a. Heysquad: A spoken question answering dataset. *arXiv preprint arXiv:2304.13689*.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023b. 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.

Chenghao Xiao, Hou Pong Chan, Hao Zhang, Weiwen Xu, Mahani Aljunied, and Yu Rong. 2025. Scaling language-centric omnimodal representation learning. *arXiv preprint arXiv:2510.11693*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

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*.

Junhan Yang, Jiahe Wan, Yichen Yao, Wei Chu, Yinghui Xu, Emma Wang, and Yuan Qi. 2025. inf-retriever-v1 (revision 5f469d7).

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

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.

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

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

This Appendix section expands on the main paper with full benchmark protocols and per-task results for all four benchmarks, the cascaded-caption data-generation pipeline (system prompts and one qualitative example per modality), per-source dataset attributions and licenses, the mining subsystem, the hyperparameter table for the smallest variant, and a short list of limitations.

A Detailed benchmark protocols

For each benchmark we briefly describe the source benchmark, the exact subset of tasks we evaluate, and the metric. We then report per-task numbers for the established public baselines and our Omni-Embed-Mini variants. Parameter counts in the left-most numeric column are taken from each model's published card; closed-source models with undisclosed sizes are marked "?".

A.1 MTEB-v2 (8-task English BEIR subset)

What it is. MTEB-v2 (Muennighoff et al., 2023) is the second release of the Massive Text Embedding Benchmark; it modernises the evaluation API and unifies multilingual (MMTEB) and image (MIEB) extensions, with the full v2 catalog spanning 1000+ task-language pairs across 250+ languages.

Subset evaluated. Even at sub-billion parameters, evaluating the full MTEB-v2 catalog requires days of GPU time, and many embedding papers report on a curated BEIR subset for the same reason. We use an 8-task English BEIR subset spanning short and long queries against short and long corpora: ArguAna, CQADupstack-English, CQADupstack-Physics, CQADupstack-Programmers, FiQA-2018, NFCorpus, SCIDOCS, SciFact.

Metric. nDCG@10 ($\times 100$) per task; Avg is the mean across the 8 tasks. Per-task results are reported in Table 3.

A.2 MAEB (22-task English subset)

What it is. MAEB (El Assadi et al., 2026) is a unified MTEB-ecosystem benchmark for audio embeddings that covers speech, environmental sound, and music with both audio-only and text-query tasks.

Subset evaluated. We evaluate a 22-task subset comprising the 15 native-English tasks plus 7 language-independent music/environmental-sound tasks; multilingual source tasks are restricted to their English split. Tasks split naturally into a 10-task *audio* subset and a 12-task *speech* subset, each reported in a separate table below. The benchmark task families, following (El Assadi et al., 2026), are: text-to-audio retrieval (T2A), classification (Cls), clustering (Clu), multi-label tagging (Mlt), pair-classification (Pair), reranking (Rank) and zero-shot classification (ZSh).

Metrics. T2A reports HitRate@5; Cls and ZSh report accuracy; Clu reports V-measure; Mlt reports LRAP; Pair reports max average precision; Rank reports MAP@1000. Avg is the mean across each model's scored tasks within the relevant modality subset. Per-task results are reported in Table 4 (audio subset) and Table 5 (speech subset).

A.3 MMEB-V2 (16-task English subset)

What it is. MMEB-V2 (Meng et al., 2025) is a multimodal embedding benchmark combining image and video tasks under a unified hit@1 retrieval and classification protocol.

Subset evaluated. We use the 16 English tasks: 10 image tasks (VOC2007, Country211, OK-VQA, EDIS, MSCOCO-T2I, VisualNews-T2I, MSCOCO-I2T, VisualNews-I2T, NIGHTS, FashionIQ) and 6 video tasks (HMDB51, UCF101, MSR-VTT, MSVD, DiDeMo, VATEX). Following the benchmark task taxonomy, image tasks span zero-shot image classification (CLS), visual question answering (VQA), text-to-image (T2I), image-to-text (I2T) and image-to-image (I2I) retrieval; video tasks span zero-shot action classification (V-CLS) and text-to-video (V-T2V) retrieval. The MMEB-V2 NExTQA and EgoSchema video-QA tasks are not included in this subset.

Metric. hit@1 ($\times 100$) per task; AvgImg, AvgVid and AvgAll are means across the image, video, and combined task sets respectively. Per-task results are reported in Table 6 (image tasks) and Table 7 (video tasks).

A.4 ViDoRe-V3 (7 English domains)

What it is. ViDoRe-V3 (Loison et al., 2026) is a visual-document retrieval benchmark over visually rich enterprise documents (financial reports, HR handbooks, scientific papers, regulatory filings, pharmaceutical and industrial manuals): given a text query, retrieve the relevant page-image.

Subset evaluated. We evaluate the 7 English-source domains (Finance, HR, Industrial, Pharmaceuticals, Computer Science, Energy, Physics); the

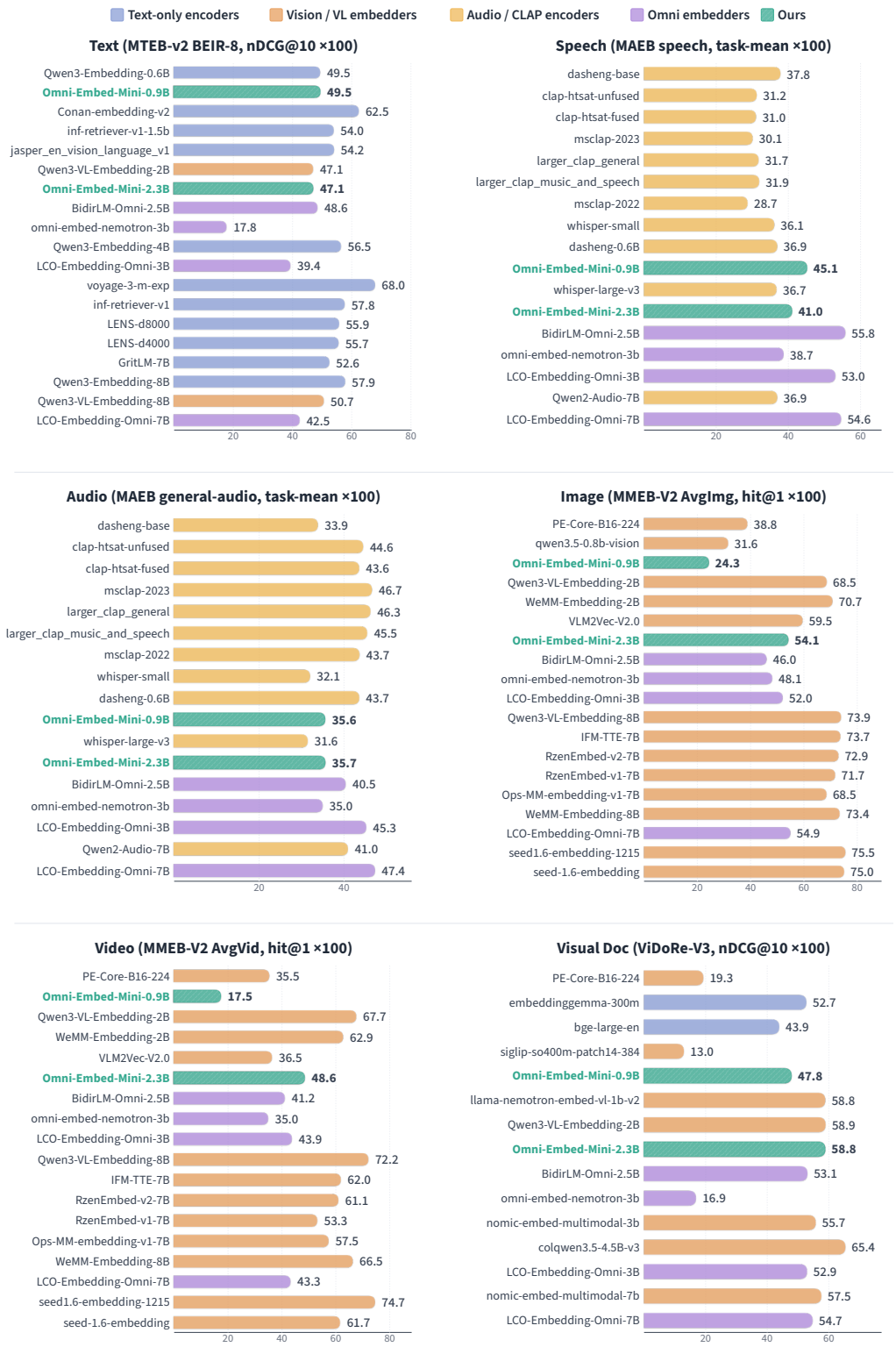


Figure 8: **Per-modality rankings, all classes.** Each panel ranks every model with a score for that modality, descending; bars are coloured by class (text-only, VL/vision, audio/CLAP, omni, ours). Text is MTEB-v2 BEIR-8 Avg; Speech/Audio are mean across the 12 speech and 10 general-audio MAEB tasks; Image/Video are MMEB-V2 AvgImg/AvgVid; Visual Doc is ViDoRe-V3 Avg nDCG@10. Omni-Embed-Mini-0.9B preserves backbone MTEB nDCG@10 (49.50, matching our standalone reproduction) while extending to five additional modalities at $\sim$2.7–9.5$\times$ fewer parameters than every omni baseline. The 2.3B variant additionally inherits Qwen3-VL-Embedding-2B image/video/visual-doc strengths.

Model	Params	ArgA	CQA-E	CQA-P	CQA-Pr	FiQA	NFC	SCID	SciF	Avg
Text embedders										
voyage-3-m-exp	6.9B	**89.79**	**69.98**	**73.71**	**65.87**	**78.01**	46.99	**34.53**	85.09	**68.00**
Conan-embedding-v2	1.5B	77.79	57.50	60.58	60.66	71.86	**55.74**	31.63	84.27	62.50
Qwen3-Embedding-8B	7.6B	76.85	57.44	58.31	53.49	64.57	41.45	32.74	78.46	57.91
inf-retriever-v1	7.1B	84.86	55.96	52.35	46.64	62.35	43.66	30.78	**85.40**	57.75
Qwen3-Embedding-4B	4.0B	75.64	55.00	56.58	51.17	62.65	41.10	31.44	78.33	56.49
LENS-d8000	7.1B	76.02	55.47	53.89	51.15	61.57	40.61	28.54	79.75	55.87
LENS-d4000	7.1B	77.32	55.62	52.87	51.53	60.43	41.64	27.51	78.39	55.66
inf-retriever-v1-1.5B	1.5B	81.50	52.16	49.68	45.19	56.09	38.55	26.34	82.84	54.04
GritLM-7B	7.2B	63.17	52.10	52.85	48.29	59.91	40.86	24.41	79.13	52.59
Qwen3-Embedding-0.6B	596M	69.70	51.45	50.61	46.32	47.23	36.72	24.27	69.70	49.50
Vision–language embedders										
Jasper-EN-VL-v1	1.6B	65.20	54.27	55.00	50.45	60.12	43.31	24.64	80.37	54.17
Qwen3-VL-Embedding-8B	8.1B	62.77	52.88	52.13	46.50	47.62	39.75	24.72	79.19	50.70
Qwen3-VL-Embedding-2B	2.1B	58.46	48.27	50.66	43.44	42.30	37.05	21.79	74.73	47.09
Omni / audio–language embedders										
BidirLM-Omni-2.5B	2.5B	57.04	49.91	**51.67**	45.93	45.60	**38.67**	23.32	**76.31**	48.56
LCO-Embedding-Omni-7B	8.9B	39.09	44.47	50.59	43.18	39.21	36.35	19.97	67.42	42.54
LCO-Embedding-Omni-3B	4.7B	41.39	39.51	45.67	40.14	35.47	31.86	18.21	63.13	39.42
omni-embed-nemotron-3B	3.0B	59.45	4.93	8.71	4.39	1.93	4.21	1.71	57.37	17.84
Omni-Embed-Mini-0.9B	935M	**69.70**	**51.45**	50.61	**46.32**	**47.23**	36.72	**24.27**	69.70	**49.50**
Omni-Embed-Small-2.3B	2.3B	58.46	48.27	50.66	43.44	42.30	37.05	21.79	74.73	47.09

Table 3: MTEB-v2 BEIR-8 per-task scores. All eight tasks measure text retrieval ranking quality, scored by nDCG@10 ($\times 100$). **Task abbreviations:** ArguAna (ArgA), CQADupstack-English (CQA-E), CQADupstack-Physics (CQA-P), CQADupstack-Programmers (CQA-Pr), FiQA-2018 (FiQA), NFCorpus (NFC), SCIDOCS (SCID), SciFact (SciF). Within the omni block, per-column best in **bold**, second best underlined; our row highlighted. Omni-Embed-Mini's text inference path is byte-identical to the frozen Qwen3-Embedding-0.6B backbone (no LoRA or projector applied to the text branch), so adding multimodal heads and dense-caption distillation does not regress text retrieval.

French-only Finance task is excluded.

Metric. nDCG@10 per task; Avg is the mean across the 7 domains. The public ViDoRe leaderboard headline reports nDCG@5; we report nDCG@10 to match the convention used elsewhere in this paper. Per-domain results are reported in Table 8.

B Cascaded-caption data generation

The caption slice of our omni-modal corpus is the sole training corpus for the contrastive objective. Every row is the output of a single multimodal call to Qwen3-Omni-30B-A3B-Instruct at temperature 0.7 with a 2048-token output budget. The model receives the raw media plus a modality-specific system prompt that interpolates the source dataset's ground-truth caption as a "Reference caption" grounding prior; a short user-facing prompt is drawn at random from a per-modality bank and concatenated with the appropriate media tag.

B.1 Modality system prompts

System prompt: Audio

Describe only what you hear; no opinions, speculation, or filler. Be precise and factual. Do not hallucinate sounds not present. Listen to this audio. Reference caption: "{original_text}". Write a dense, semantically rich description. Cover: each distinct sound and its characteristics (pitch, volume, texture); temporal order (what happens first, next, last); acoustic environment (indoor/outdoor, reverb, background); patterns or repetitions. Be concise; every word should convey information about the audio. One paragraph.

System prompt: Music (audio music subsets)

Describe only what you hear. Listen to this music. Reference caption: "{original_text}". Write a dense description covering: instruments and timbres; melody, harmony, key, chord progressions; rhythm, tempo (BPM estimate), time signature; genre, energy, dynamics; structure; vocals if present (gender, style, lyrics gist). One paragraph.

Model	Params	Clotho HR@5	MACS HR@5	US8K HR@5	BjOp. Acc	Bird Acc	GTZAN Acc	Mrd. Acc	Veh. V-meas.	FSD19 LRAP	GTZ-R MAP@1k	Avg
Audio–language models												
Qwen2-Audio-7B	7.0B	0.45	1.53	0.22	97.45	37.10	93.10	61.17	5.52	32.64	80.85	41.00
CLAP-style dual encoders												
CLAP-LAION (general)	194M	33.72	33.08	0.96	93.62	17.00	84.50	49.45	3.37	80.63	66.78	46.31
CLAP-LAION (music+speech)	194M	32.39	30.53	0.94	91.52	16.40	83.50	48.99	4.72	80.62	65.65	45.53
MS-CLAP 2023	160M	41.93	27.23	0.94	91.11	17.30	78.10	52.09	2.86	79.73	75.43	46.67
MS-CLAP 2022	196M	58.73	41.73	0.88	93.20	13.20	58.70	29.14	7.98	70.86	62.91	43.73
CLAP-HTSAT (unfused)	153M	34.18	27.99	0.98	91.96	16.40	74.90	49.52	2.66	80.84	66.45	44.59
CLAP-HTSAT (fused)	154M	33.57	29.26	0.94	92.77	18.00	67.90	47.00	4.71	81.09	61.30	43.65
Speech encoders (no text tower; avg over 7 non-retrieval columns†)												
Dasheng-0.6B	630M	–	–	–	84.33	8.50	76.20	53.83	5.76	7.12	69.98	43.67†
Dasheng-base	85M	–	–	–	76.73	5.10	65.40	44.50	3.22	7.08	35.56	33.94†
Whisper-large-v3	1.5B	–	–	–	64.80	5.50	45.90	30.80	1.49	8.59	63.77	31.55†
Whisper-small	242M	–	–	–	67.38	7.30	47.40	27.20	0.98	10.29	64.16	32.10†
Omni embedders												
LCO-Embedding-Omni-7B	8.9B	**33.89**	**29.77**	**0.92**	92.79	**34.10**	**82.30**	42.07	3.63	**75.74**	**78.71**	**47.39**
LCO-Embedding-Omni-3B	4.7B	32.70	22.65	0.88	**93.22**	31.60	81.00	39.54	2.18	74.05	75.37	45.32
BidirLM-Omni-2.5B	2.5B	12.05	8.40	0.65	90.25	33.10	74.10	**45.35**	5.11	60.95	74.76	40.47
omni-embed-nemotron-3B	3.0B	10.40	7.12	0.59	83.92	15.70	63.80	28.97	0.90	72.56	66.52	35.05
Ours: Omni-Embed-Mini												
0.9B, no mining	935M	17.94	8.65	0.75	81.77	16.60	65.80	26.89	4.07	50.92	76.20	34.96
0.9B, text mining	935M	13.45	6.11	0.63	84.32	17.70	66.30	37.27	7.68	47.53	70.90	35.19
0.9B, media mining	935M	19.07	8.14	0.73	83.88	18.00	69.10	34.73	2.77	51.09	76.49	36.40
0.9B, text + media	935M	16.42	10.94	0.86	82.64	13.50	66.10	31.06	**7.95**	51.05	75.90	35.64
0.9B, t+m, orig. captions	935M	21.07	11.96	0.86	73.72	11.60	50.10	36.98	7.18	44.54	64.48	32.25
2.3B, text + media	2.3B	27.41	10.43	0.83	78.40	16.70	66.50	27.26	2.64	51.24	75.31	35.67

Table 4: MAEB-Audio per-task scores across 10 tasks. **Task abbreviations:** Clotho-T2A (Clotho), MACS-T2A (MACS), UrbanSound8K-T2A (US8K), BeijingOpera (BjOp.), BirdCLEF (Bird), GTZAN-Genre (GTZAN), MridinghamTonic (Mrd.), VehicleSoundClustering (Veh.), FSD2019Kaggle (FSD19), GTZAN-AudioReranking (GTZ-R). **Metrics** appear below each task name: HR@5 = HitRate@5, Acc = accuracy, V-meas. = V-measure, LRAP = label ranking avg. precision, MAP@1k = mean avg. precision at 1000. Within the omni block, per-column best in **bold**, second best underlined; our rows highlighted. †Speech-only encoders lack a text tower and are not scored on the three retrieval tasks; their average is over the remaining seven columns.

System prompt: Speech

Listen to this speech audio. Write a single flowing paragraph covering: (1) the speaker (age group, gender, voice pitch/tone, clarity, accent); (2) speaking style (pace, loudness, emotion); (3) a transition into the transcription using a casual phrase followed by the marker [TRANSCRIPTION]; (4) audio quality, recording environment, background sounds. Use the literal text [TRANSCRIPTION]; do not write out the actual words spoken. The placeholder is post-processed into a transcription span around the source transcript.

System prompt: Image

Look at this image. Reference caption: "{original_text}". Write a dense, factual description covering: subjects and objects (what, how many, attributes); spatial layout and composition; colors, lighting, and style; any actions or interactions; any text visible in the image (OCR). Be concrete; avoid speculation about meaning or backstory. One paragraph.

System prompt: Video

Watch this video. Reference caption: "{original_text}". Write a dense description covering: subjects and setting; spatial layout; the temporal sequence of events (what happens first, next, last); camera motion or shot transitions; on-screen actions; any text or captions visible on screen. One paragraph.

System prompt: Visual document

You are an expert document analyst. Examine this document image and provide a comprehensive description covering: document type and purpose; overall layout and structure; all text content (headings, body, labels); tables (column headers, row structure, key data); charts and graphs (type, axes, trends); figures and diagrams; visual formatting (colors, fonts, emphasis). Be thorough, precise, and well-organised.

B.2 User-prompt banks

For each modality the user-side message is sampled from a small bank of natural-language requests: e.g. "what do you hear", "describe this audio", "caption this" for audio; "what did they say", "transcribe

Model	Params	GigaSp HR@5	SpSQuAD HR@5	CREMA Acc	AgeDet Acc	IEMOCAP Acc	VoxCel Acc	CREMA-C V-meas.	CREMA-P max-AP	NMSQA max-AP	VoxPop max-AP	Ravd ZS Acc	SpCmd ZS Acc	Avg
Audio–language models														
Qwen2-Audio-7B	7.0B	0.09	1.00	73.99	17.59	92.96	29.54	32.37	68.87	48.85	52.93	14.37	10.38	36.91
CLAP-style dual encoders														
CLAP-LAION (general)	194M	0.16	2.00	39.83	20.51	89.28	30.47	13.17	54.71	47.42	53.55	17.29	12.44	31.74
CLAP-LAION (music+speech)	194M	0.21	2.00	40.18	16.59	93.12	33.78	14.43	56.66	46.14	53.49	17.29	9.03	31.91
MS-CLAP 2023	160M	0.10	0.00	37.06	15.35	85.86	26.18	10.73	57.65	51.09	52.05	15.21	10.01	30.11
MS-CLAP 2022	196M	0.16	2.33	27.94	14.27	89.35	25.54	5.36	53.06	50.09	52.08	13.61	10.51	28.69
CLAP-HTSAT (unfused)	153M	0.36	1.33	37.56	15.98	92.58	33.72	12.78	56.91	46.98	53.96	13.26	9.35	31.23
CLAP-HTSAT (fused)	154M	0.18	2.00	38.75	15.27	93.62	29.29	10.82	54.95	47.85	53.18	14.65	11.76	31.03
Speech encoders (no text tower; avg over the 8 non-retrieval / non-zero-shot columns†)														
Dasheng-0.6B	630M	–	–	39.76	20.09	57.73	16.24	8.65	55.92	46.29	50.70	–	–	36.92†
Dasheng-base	85M	–	–	39.04	17.22	53.91	24.10	9.13	55.94	52.38	50.70	–	–	37.80†
Whisper-large-v3	1.5B	–	–	41.31	17.22	50.85	16.64	9.68	57.83	48.49	51.71	–	–	36.72†
Whisper-small	242M	–	–	37.23	19.53	52.18	13.19	10.30	58.40	45.61	52.37	–	–	36.10†
Omni embedders														
LCO-Embedding-Omni-7B	8.9B	**83.29**	74.00	**36.05**	16.16	70.75	43.40	0.62	**53.56**	**97.60**	50.73	31.67	97.42	54.60
LCO-Embedding-Omni-3B	4.7B	83.13	72.00	31.03	16.87	59.42	48.97	0.62	52.40	96.92	50.67	26.94	97.40	53.03
BidirLM-Omni-2.5B	2.5B	83.05	**74.33**	29.92	15.04	**85.38**	**49.49**	0.60	53.20	95.99	51.05	**34.24**	97.30	**55.80**
omni-embed-nemotron-3B	3.0B	38.31	4.33	22.37	16.64	60.45	33.05	0.71	52.77	57.26	50.87	29.72	**97.67**	38.68
Ours: Omni-Embed-Mini														
0.9B, no mining	935M	76.02	66.67	19.13	17.15	64.64	38.24	0.55	51.68	67.18	50.67	19.86	59.65	44.29
0.9B, text mining	935M	76.09	62.33	19.66	15.19	58.30	37.20	0.57	52.05	71.17	50.38	20.07	68.11	44.26
0.9B, media mining	935M	76.34	67.67	19.40	16.07	66.62	37.84	0.57	51.94	69.20	49.60	19.65	64.04	44.91
0.9B, text + media	935M	79.73	66.67	19.00	16.85	63.35	40.45	0.62	51.81	70.80	49.96	20.42	62.10	45.15
0.9B, t+m, orig. captions	935M	78.24	65.67	18.99	**17.60**	58.12	39.29	0.59	51.33	71.33	50.32	15.62	54.32	43.45
2.3B, text + media	2.3B	68.15	50.67	20.25	17.09	64.74	36.73	**0.79**	52.48	63.02	**51.62**	14.72	51.74	41.00

Table 5: MAEB Speech per-task scores across 12 tasks. **Task abbreviations:** GigaSpeechT2A (GigaSp), SpokenSQuADT2A (SpSQuAD), CREMA-D (CREMA), CommonLanguageAgeDetection (AgeDet), IEMOCAPGender (IEMOCAP), VoxCelebSA (VoxCel), CREMA-D-Clustering (CREMA-C), CREMA-D-PairClassification (CREMA-P), NMSQA-PairClassification (NMSQA), VoxPopuli-AccentPairClassification (VoxPop), Ravdess-Zeroshot (Ravd), SpeechCommands-Zeroshot-v0.02 (SpCmd). **Metrics** appear below each task name: HR@5 = HitRate@5, Acc = accuracy, V-meas. = V-measure, max-AP = pair-classification max average precision, ZS Acc = zero-shot accuracy. Within the omni block, per-column best in **bold**, second best underlined; our rows highlighted. †Speech-only encoders lack a text tower and are not scored on the two retrieval tasks or the two zero-shot tasks; their average is over the remaining eight columns.

this" for speech; "what do you see", "describe this image" for image; and "what do you see", "describe this video", "what is happening in this video" for video. Visual-document rows use a single short user prompt; the grounding work is carried by the system prompt.

B.3 Sample schema

Each row of the caption slice carries a unique row identifier, the raw media bytes, a chat-style messages field holding the user request and the model's dense caption, the modality label (audio, speech, image, video, or visual-doc), the caption category (always "caption" for this slice), the source dataset's original short caption (used as the grounding prior), and metadata fields for the source dataset name, sampling rate (16,000 for audio/speech) and frame rate (for video).

B.4 Qualitative dense-caption examples

Table 9 shows one representative example per modality from the caption slice. The middle column is a brief written gloss of the underlying media (a one-line description sitting between the original short caption and the dense caption); the third column is the original short caption from the source dataset (the grounding prior); the fourth column is the dense caption produced by the cascaded captioner and used as the contrastive teacher target.

C Per-source datasets, licenses, citations

The caption slice draws from 16 unique source datasets across 5 modalities, totalling 242,080 caption rows and 54.9 days of media. The aggregated corpus is released under CC BY-SA 4.0; individual source datasets retain their respective licenses, summarised in Table 10.

D Mining subsystem

The miner runs as a single background thread that shares the live training model with the main loop, avoiding the roughly 16 GB extra HBM cost of a separate model copy. Within-cycle weight drift between miner and main thread is empirically negligible because the LoRA delta per optimiser step

Model	Params	VOC07 CLS	Ctry211 CLS	OKVQA VQA	EDIS T2I	COCO-T2I T2I	VN-T2I T2I	COCO-I2T I2T	VN-I2T I2T	NIGHTS I2I	FashIQ I2I	Avg
Vision–language embedders												
seed1.6-embedding-1215	?	89.30	46.90	74.90	96.70	81.30	83.80	77.80	83.70	70.70	50.40	75.55
seed-1.6-embedding	?	91.90	47.70	74.20	91.50	78.80	83.40	77.40	84.50	72.10	49.00	75.05
Qwen3-VL-Embedding-8B	8.1B	93.40	27.30	77.80	96.30	81.10	81.10	79.10	85.70	72.70	44.50	73.90
IFM-TTE-7B	8.3B	83.40	60.50	83.40	95.50	77.70	79.20	72.60	82.50	69.40	32.40	73.66
WeMM-Embedding-8B	8.8B	94.50	32.20	73.40	94.90	81.20	81.70	78.20	84.90	69.00	43.80	73.38
RzenEmbed-v2-7B	8.3B	91.40	29.70	73.10	97.20	81.10	82.60	78.00	85.80	69.00	41.50	72.94
RzenEmbed-v1-7B	8.3B	91.60	29.00	69.60	96.50	78.60	82.60	76.00	85.80	68.60	38.70	71.70
WeMM-Embedding-2B	2.1B	93.70	28.90	66.40	93.30	77.60	76.90	77.00	80.30	69.60	43.20	70.69
Qwen3-VL-Embedding-2B	2.1B	92.70	24.40	66.20	92.50	77.30	73.40	74.60	75.90	69.50	38.80	68.53
Ops-MM-embedding-v1-7B	8.3B	84.80	28.50	70.60	94.40	79.30	80.10	72.10	84.30	66.20	24.30	68.46
VLM2Vec-V2.0	2.2B	76.80	19.50	44.00	82.90	73.80	72.30	66.40	73.70	66.20	19.30	59.49
qwen3.5-0.8b-vision (I2I only)	100M	–	–	–	–	–	–	–	–	59.30	3.80	31.55†
Omni embedders												
LCO-Embedding-Omni-7B	8.9B	56.40	**33.20**	**59.90**	66.70	64.60	62.40	61.20	61.20	62.90	20.70	**54.92**
LCO-Embedding-Omni-3B	4.7B	**71.20**	28.90	58.20	56.50	59.50	55.20	59.40	59.20	56.20	15.90	52.02
omni-embed-nemotron-3B	3.0B	53.30	23.40	16.30	82.60	60.40	60.70	52.20	59.50	63.80	8.70	48.09
BidirLM-Omni-2.5B	2.5B	68.10	8.00	20.30	72.90	**72.10**	48.50	63.60	30.60	62.60	13.10	45.98
Ours: Omni-Embed-Mini												
0.9B, no mining	935M	64.10	2.20	6.50	44.60	31.60	9.90	18.70	3.90	41.50	2.50	22.55
0.9B, text mining	935M	57.60	1.30	4.70	42.50	27.20	6.40	10.80	4.30	40.10	1.90	19.68
0.9B, media mining	935M	61.60	1.50	5.70	43.60	34.70	12.30	20.10	6.10	55.40	4.70	24.57
0.9B, text + media	935M	48.00	2.30	11.00	47.10	38.00	14.00	20.70	7.90	50.00	4.10	24.31
0.9B, t+m, orig. captions	935M	61.00	1.30	1.30	35.40	15.00	5.80	5.70	3.30	45.30	2.50	17.66
2.3B, text + media	2.3B	21.10	18.40	41.30	**84.60**	70.50	**65.50**	**68.50**	**68.50**	**68.00**	**34.50**	54.09

Table 6: MMEB-V2 image per-task hit@1 ($\times 100$). **Task abbreviations:** VOC2007 (VOC07), Country211 (Ctry211), OK-VQA (OKVQA), EDIS, MSCOCO-T2I (COCO-T2I), VisualNews-T2I (VN-T2I), MSCOCO-I2T (COCO-I2T), VisualNews-I2T (VN-I2T), NIGHTS, FashionIQ (FashIQ). **Task families** appear below each task name: CLS = zero-shot classification accuracy; VQA = knowledge retrieval hit@1; T2I = text-to-image retrieval hit@1; I2T = image-to-text retrieval hit@1; I2I = image-to-image retrieval hit@1. Within the omni block, per-column best in **bold**, second best underlined; our rows highlighted. †qwen3.5-0.8b-vision supports only the two I2I tasks; its average is over those two columns and is not directly comparable to the others.

is small and the embedding geometry is dominated by the frozen backbone.

Cadence. The miner alternates between two cycle types gated on the training step counter. A *text* cycle (re-embedding captions, building a CPU FAISS index, and writing hard-negative ids into the negative cache) fires roughly ten times per epoch and finishes in about thirty seconds. A *media* cycle covers all four media modalities (audio, speech, image, video) and fires roughly five times per epoch, taking 12 to 20 minutes per cycle. Triggers are coalesced so that multiple firings during a long cycle collapse to a single follow-up.

Concurrency and cache semantics. A single forward-lock serialises miner and main-thread forward passes at batch granularity: the main thread holds the lock for the entire forward, backward and optimiser step, and the miner acquires it per encode-batch, yielding fine-grained alternation rather than overlap. The negative cache is keyed by sample id and overwrites on write, so media-mined entries supersede earlier text-mined ones for the same sample. If the cache is empty for a given sample at batch time the collator falls back to in-batch negatives plus distillation, recovering the no-mining ablation behaviour with no special-case code.

E Additional ablations

Cross-modal alignment probe. During training we monitor a SciFact text-vs-speech NDCG@10 convergence probe every 10% of an epoch. The synthesised speech version of each claim is encoded through the full multimodal path (Whisper + Dasheng $\rightarrow$ projectors $\rightarrow$ backbone) while the corpus remains text. The two NDCG@10 curves converging is the headline indicator that the audio path has reached the backbone's text geometry.

Emergent media-to-media alignment. A direct consequence of routing every modality through the *same* shared backbone and aligning each one to the caption text is that, after training, all media live in one geometry. We never train a media-to-media objective, yet image$\leftrightarrow$audio, audio$\leftrightarrow$video, image$\leftrightarrow$document and similar pairings are usable

Model	Params	HMDB V-CLS	UCF V-CLS	MSRVTT V-T2V	MSVD V-T2V	DiDeMo V-T2V	VATEX V-T2V	Avg
Vision–language embedders								
seed1.6-embedding-1215	?	93.50	98.70	60.20	74.48	66.43	54.87	74.70
Qwen3-VL-Embedding-8B	8.1B	83.40	95.10	58.20	75.67	66.04	54.87	72.21
Qwen3-VL-Embedding-2B	2.1B	78.70	91.80	53.80	75.07	57.87	49.11	67.73
WeMM-Embedding-8B	8.8B	61.50	84.20	55.30	73.43	69.22	55.05	66.45
WeMM-Embedding-2B	2.1B	57.70	75.10	53.10	73.43	65.24	52.52	62.85
IFM-TTE-7B	8.3B	65.40	79.60	52.70	73.13	49.70	51.45	62.00
seed-1.6-embedding	?	63.90	74.20	55.30	71.34	56.67	48.77	61.70
RzenEmbed-v2-7B	8.3B	63.40	75.70	50.60	72.09	57.27	47.57	61.10
Ops-MM-embedding-v1-7B	8.3B	59.30	78.80	49.40	67.16	46.12	43.95	57.45
RzenEmbed-v1-7B	8.3B	56.70	73.10	45.10	64.93	41.24	38.57	53.27
VLM2Vec-V2.0	2.2B	35.20	57.30	28.80	43.73	29.98	23.83	36.47
Omni embedders								
LCO-Embedding-Omni-3B	4.7B	39.00	65.10	38.40	58.06	31.47	31.13	43.86
LCO-Embedding-Omni-7B	8.9B	30.50	64.50	**41.50**	59.10	32.37	32.05	43.34
omni-embed-nemotron-3B	3.0B	22.70	36.50	35.80	55.52	32.67	27.11	35.05
BidirLM-Omni-2.5B†	2.5B	–	–	–	–	–	–	–
Ours: Omni-Embed-Mini								
0.9B, no mining	935M	6.80	16.00	12.90	16.87	7.77	5.07	10.90
0.9B, text mining	935M	8.30	18.70	10.80	13.13	6.08	3.66	10.11
0.9B, media mining	935M	11.20	24.50	15.40	20.90	9.46	6.05	14.59
0.9B, text + media	935M	12.60	25.30	18.20	26.87	13.35	8.80	17.52
0.9B, t+m, orig. captions	935M	6.20	6.70	6.60	7.16	2.29	1.85	5.13
2.3B, text + media	2.3B	**44.20**	**74.30**	40.40	**61.04**	**37.45**	**34.35**	**48.62**

Table 7: MMEB-V2 video per-task hit@1 ($\times 100$). **Task abbreviations:** HMDB51 (HMDB), UCF101 (UCF), MSR-VTT (MSRVTT), MSVD, DiDeMo, VATEX. **Task families** appear below each task name: V-CLS = zero-shot action classification accuracy; V-T2V = text-to-video retrieval hit@1. Across all omni-style models, the 8 frames per video are embedded as images and mean-pooled to keep the comparison uniform. Within the omni block, per-column best in **bold**, second best underlined; our rows highlighted. †BidirLM-Omni-2.5B does not declare a video modality and is not scored on these tasks; the row is listed for completeness.

at inference because they share the text axis as a common reference frame. This is the same emergence ImageBind (Girdhar et al., 2023) reports, but obtained without contrasting modalities pair-wise: text-as-anchor implicitly induces a fully-connected graph, and the frozen text-embedding backbone gives that graph a well-calibrated metric for free.

Single-stage training. LLaVA (Liu et al., 2023) and most VL models train in two distinct stages with two distinct datasets. Omni-Embed-Mini runs as a *single-stage* pipeline: the dataset is loaded once and the only "phasing" is internal to the optimisation schedule (projector-only → LoRA adapters → online miner activation). The two phases share the same data-loader, the same dense cascaded captions, and the same loss, with no hand-off. This is possible because dense cascaded captions carry both alignment signal and task signal in one row.

F Loss equations

We give the full form of the Matryoshka SigLIP contrastive loss summarised in Sec. 3.4. Let $\hat{a}_i^{(d)}$, $\hat{p}_j^{(d)}$, $\hat{h}_k^{(d)}$ denote the L2-normalised, d-truncated student, in-batch positive, and hard-negative embeddings. With learned temperature $\tau = \exp(\log \tau)$ (clamped to $[0.01, 0.3]$) and learned bias b,

$$z_{ij}^{(d)} = \frac{1}{\tau}\, \hat{a}_i^{(d)} \cdot \hat{p}_j^{(d)} + b, \quad y_{ij} = \begin{cases} +1 & i = j \\ -1 & i \neq j \end{cases}, \tag{1}$$

$$\zeta_{ik}^{(d)} = \frac{1}{\tau}\, \hat{a}_i^{(d)} \cdot \hat{h}_k^{(d)} + b, \tag{2}$$

Model	Params	Fin	HR	Ind	Phr	CS	En	Phy	Avg
Visual-document retrievers									
colqwen3.5-4.5B-v3	4.6B	68.69	66.09	59.35	67.31	79.87	66.16	50.08	65.36
nomic-embed-multimodal-7b	7.8B	59.43	58.27	48.60	63.44	71.47	56.96	44.70	57.55
nomic-embed-multimodal-3b	3.8B	57.84	54.88	46.63	60.35	70.29	54.96	45.17	55.73
Vision–language embedders									
Qwen3-VL-Embedding-2B	2.1B	63.68	58.70	47.55	64.60	74.41	56.80	46.74	58.93
llama-nemotron-embed-vl-1b-v2	1.7B	64.53	60.59	47.85	63.52	74.01	57.72	43.65	58.84
google/siglip-so400m-p14-384	878M	11.04	11.22	8.91	22.14	19.58	7.57	10.25	12.96
Text retrievers (operate on OCR'd page text)									
embeddinggemma-300m	308M	51.03	49.06	42.44	60.28	67.07	55.74	43.21	52.69
BAAI/bge-large-en	335M	38.71	45.71	39.39	55.81	64.44	30.48	32.58	43.88
Omni embedders									
LCO-Embedding-Omni-7B	8.9B	55.13	55.31	41.94	61.14	69.43	55.44	44.64	54.72
BidirLM-Omni-2.5B	2.5B	58.87	51.58	41.17	56.23	68.85	51.92	42.77	53.06
LCO-Embedding-Omni-3B	4.7B	54.40	52.06	39.37	58.99	67.38	53.86	44.31	52.91
omni-embed-nemotron-3b	3.0B	14.17	17.97	10.96	18.54	23.45	8.81	24.19	16.87
Ours: Omni-Embed-Mini									
0.9B, no mining	935M	8.62	5.24	5.62	10.55	9.76	5.41	3.65	6.98
0.9B, text mining	935M	2.84	2.33	2.71	1.84	3.98	1.99	1.47	2.45
0.9B, media mining	935M	3.29	5.39	1.86	1.36	6.57	2.78	1.29	3.22
0.9B, text + media	935M	49.50	43.93	38.86	51.92	64.68	48.40	37.56	47.84
0.9B, t+m, orig. captions	935M	0.38	1.16	0.21	0.25	0.38	0.23	0.77	0.48
2.3B, text + media	2.3B	**63.19**	**58.99**	**47.57**	**64.39**	**74.34**	**56.95**	**46.17**	**58.80**

Table 8: ViDoRe-V3 per-domain scores. All seven tasks are text-to-page retrieval (T2Pg) over visual-document page images, scored by nDCG@10 ($\times 100$). **Domain abbreviations:** Finance (Fin), HR, Industrial (Ind), Pharmaceuticals (Phr), Computer Science (CS), Energy (En), Physics (Phy). Within the omni block, per-column best in **bold**, second best underlined; our rows highlighted. The mining ablation rows show that visual-document retrieval is the modality where hybrid hard-negative mining is most decisive: text-only and media-only mining barely move the projector off the prior, but combined mining crosses 47 nDCG@10 on the 0.9B model, a $\sim 7\times$ swing relative to either single-side ablation.

and the per-dim SigLIP loss is the mean over all concatenated pairs:

$$\mathcal{L}_{\text{siglip}}^{(d)} = -\frac{1}{B(B+K)}\left[\sum_{i,j}\log\sigma(y_{ij}z_{ij}^{(d)}) + \sum_{i,k}\log\sigma(-\zeta_{ik}^{(d)})\right] \tag{3}$$

The Matryoshka aggregation uses $1/\sqrt{d}$ weighting, $w_d = (1/\sqrt{d})/\sum_{d'}(1/\sqrt{d'})$, upweighting aggressive truncations to reflect their higher per-coordinate retrieval cost. The cosine distillation term is $\mathcal{L}_{\text{dist}} = \sum_d w_d \mathbb{E}_i[1 - \hat{s}_i^{(d)} \cdot \hat{t}_i^{(d)}]$, and the total loss is $\mathcal{L} = \mathcal{L}_{\text{contr}} + \lambda\,\mathcal{L}_{\text{dist}}$ with $\lambda = 1$.

G Hyperparameters and training schedule

Table 11 lists the hyperparameters of the smallest variant, Omni-Embed-Mini-0.9B. Training runs on 8 GPUs with DDP, BF16 mixed precision, AdamW, and a single cosine schedule with warmup. Batch size is held near-uniform across modalities (image, visual document, audio and speech at 32; video at 8 due to its patch-token footprint).

H Limitations

- **Vision modes diverge between the 0.9B and 2.3B variants.** The 0.9B variant uses a ViT extracted from the vision-native Qwen3.5-0.8B model because the 0.6B text backbone has no

vision module; the 2.3B variant reuses its backbone's native vision module. The training configuration accommodates both, but cross-variant ablations of the two vision strategies are left to future work.

- **Visual-document captions are roughly six times longer than other modalities.** Visual-document dense captions average about 4,157 characters versus about 582 for audio. The frozen backbone's 2,048-token context truncates the long tail of visual-document captions; truncation is consistent between teacher and student paths so cosine alignment is unaffected, but the upper end of visual-doc detail is lost from the contrastive signal.

- **Video uses per-frame mean-pooled embeddings on MMEB-V2.** Across all omni-style models we evaluate, the 8 frames per video are embedded as images and mean-pooled. This is uniform across the comparison and keeps results directly comparable, but caps achievable video performance below what a native temporal encoder could in principle reach.

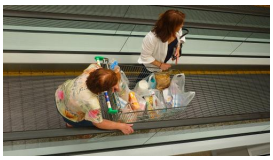
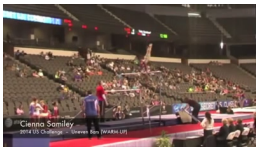
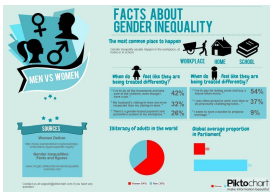
Modality	Media (described)	Original caption	Dense caption
Audio	A woman speaks in a public space while a slow heartbeat-like thump and faint clock-like ticking play in the background.	A woman talks while a heartbeat and ticking play in the background.	A woman's voice, speaking in a mid-to-low pitch, is present at the beginning of the clip, accompanied by faint, indistinct conversation from other people. Simultaneously, a deep, rhythmic thumping with a slow, steady tempo, resembling a heartbeat, begins and continues throughout. The background includes a persistent, low-level ambient hum and the general murmur of a public indoor space with some reverberation.
Speech	A young woman with a clear, neutral accent reads a short factual question in a calm, measured voice in a quiet indoor setting.	What are organic plants understood to be?	A young adult female with a high-pitched, clear, and articulate voice delivers the line in a standard, neutral accent. Her speaking style is measured and calm, with a moderate pace and consistent volume, conveying a neutral, informative affect. She says, "What are organic plants understood to be?" The audio is clean and crisp, likely recorded in a quiet indoor space with no discernible background noise or reverberation.
Image		Two women going down a conveyor belt with a cart full of groceries.	The image, taken from a high-angle perspective, shows two women on a descending escalator. The woman in the foreground, wearing a floral-patterned short-sleeved shirt, leans forward, gripping the handle of a metal shopping cart. Behind her, another woman with brown hair, wearing a white top over a dark shirt, also holds the cart's handle. The cart is filled with various items, including a visible bag of potato chips and other grocery products. The setting appears to be indoors, with a neutral-toned, geometric-patterned wall in the background; lighting is even and diffuse, characteristic of an indoor public space.
Video		There is a girl gymnast who is performing on the bars.	A gymnast, Cienna Samiley, performs on the uneven bars during the 2014 US Challenge warm-up. The scene is a large, indoor arena with sparse audience seating. The camera cuts to a podium where seven young female gymnasts stand, wearing leotards and medals, in front of a banner reading "2014 SECRET U.S. CLASSIC". A scoreboard graphic is superimposed, displaying the top six scores. The gymnasts on the podium then raise their arms and cheer.
Visual doc		(not available)	A detailed analysis of an infographic titled "Facts About Gender Inequality," which uses a combination of text, icons, and data visualizations to present information on gender disparities. The infographic is set against a light teal background, structured into several sections, with a bold sans-serif title at the top … Icons use simple black silhouettes (gender symbols, a wheelbarrow for the workplace, a house for home, books for school); the colour scheme is teal/red/black; data is presented via a pie chart and a bar chart with a consistent red-for-women, blue-for-men encoding.

Table 9: One sample per modality from the caption slice. Visual documents in the source corpus carry no human-written original caption; the cascaded captioner generates the dense description directly from the page image, and we mark such cells "(not available)". Visual-document dense captions are themselves substantially longer and richer than the other modalities; the example shown here is truncated for layout (indicated by "…").

Source	Modality	Samples	License
TED-LIUM 3 (Hernandez et al., 2018)	speech	26,826	CC BY-NC-ND 3.0
HeySQuAD (Wu et al., 2023a)	speech	7,199	CC BY 4.0
Spoken Alpaca (Shih et al., 2023)	speech	5,135	CC BY-NC 4.0
Gemini Speech (SB, 2025)	speech	4,726	Apache 2.0
LibriSpeech (Panayotov et al., 2015)	speech	2,854	CC BY 4.0
WavCaps (FreeSound) (Mei et al., 2024)	audio	26,217	Academic / research only
WavCaps (AudioSet-SL) (Mei et al., 2024)	audio	10,831	Academic / research only
AudioCaps (Kim et al., 2019)	audio	4,518	MIT (academic only)
WavCaps (BBC) (Mei et al., 2024)	audio	3,115	Academic / research only
Clotho (Drossos et al., 2020)	audio	394	Tampere Univ. license
WavCaps (SoundBible) (Mei et al., 2024)	audio	118	Academic / research only
ShareGPT-4o (image) (Chen et al., 2024)	image	42,271	MIT (academic only)
Recap-COCO-30K (Li et al., 2024)	image	22,474	CC BY 4.0
Flickr30k (Young et al., 2014)	image	21,398	Research-only (Flickr ToU)
ActivityNet (Krishna et al., 2017)	video	27,526	CC BY 4.0 (annotations); YouTube ToS
MSR-VTT (Xu et al., 2016)	video	6,612	Research-only (Microsoft)
ShareGPT-4o (video) (Chen et al., 2024)	video	1,555	MIT (academic only)
VisRAG (Yu et al., 2024)	visual doc	14,995	Mixed (per sub-source)
ColPali (Faysse et al., 2025)	visual doc	13,316	Mixed (per sub-source)

Table 10: Per-source breakdown of the caption slice of our omni-modal corpus, with citations and licenses. Aggregated release license is CC BY-SA 4.0.

Component	Setting
Architecture	
Backbone, frozen	Qwen3-Embedding-0.6B
Hidden size	1024
MRL dimensions	$\{128, 256, 512, 1024\}$
Vision encoder	ViT extracted from Qwen3.5-0.8B
Image resolution	224×224
Video tokens / max frames	196 / 96
Audio encoders, 0.9B	Whisper-small encoder + Dasheng-base
Audio encoders, 2.3B	Whisper-large-v3 encoder + Dasheng-0.6B
Tokens per audio encoder	128
Max audio duration	60 s
Optimisation	
LoRA rank / α / dropout	16 / 32 / 0.0
LR, audio adapters	2×10^{-4}
LR, vision adapters	1×10^{-4}
LR, encoder LoRA	1×10^{-4}
Optimizer	AdamW, weight decay 0.01
Warmup	2% of total steps
Max gradient norm	1.0
Max sequence length	2048
Batch size, image / visual-doc / audio / speech / video	32 / 32 / 32 / 32 / 8
Epochs	2, for reported runs
Mixed precision	BF16
GPUs	8, DDP
Loss and schedule	
Distillation loss weight	1.0, cosine
Initial temperature	0.07, learnable
Phase-2 switch	20% of total steps
Hard-negative mining	
Text-mining refresh	$10\times$ per epoch
Media-mining refresh	$5\times$ per epoch
Text index size	50,000 captions
Media index size	10,000 per modality
Hard negatives per sample	5
False-negative threshold	0.92, cosine

Table 11: Hyperparameters for the Omni-Embed-Mini-0.9B variant. The 2.3B variant differs in backbone, using Qwen3-VL-Embedding-2B with native vision, and in audio encoder scale, using Whisper-large-v3 with Dasheng-0.6B. Proportional batch sizes are adjusted, while optimisation settings remain unchanged.