OMNIROUTER: ROUTING, NOT SCALE A MULTIMODAL RETRIEVAL ENCODER

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal retrieval requires a single space in which text, images, ambient audio and speech can be compared, and the dominant way to obtain one is to train a unified tower across all modalities. We ask whether such a space can instead be *constructed* from specialist encoders that were never trained together, and what the construction gains and loses. Two obstacles make this hard: the geometries of independently pretrained encoders are not compatible, so choosing any one of them as the shared space discards what the others capture, and practical systems truncate embeddings, which removes an appended second geometry entirely. OMNIROUTER addresses both. A closed-form ridge mapping between two text spaces defines a residual that holds what a vision-aligned encoder knows and a text encoder does not, and the two geometries are interleaved coordinate-wise so that every truncated prefix retains both. Because each modality reaches the space through its own projector, similarity decomposes per block and a change to one route leaves the scores of directions that do not involve it exactly unchanged. Both geometries contribute on every cross-modal direction, by up to 0.149 NDCG@10 over the better single one, and interleaving beats concatenation from 256 to 1024 dimensions. The system covers four input families with 23.5M trained parameters, 1.34% of a 1.76B model, and exceeds an 8.9B omni-modal embedder on all five core directions. Explicit routing also makes failures diagnosable: a speech family that looked like a capability limit was a missing dispatch rule, and one routing change raises its mean NDCG@10 from 0.0109 to 0.7657 with no retraining. Adding audio to a text-and-image system costs 150 seconds and changes no existing embedding, whereas adapting a jointly trained tower to the same data moves its text and image embeddings.

1 INTRODUCTION

Multimodal retrieval requires a single space in which text, images, ambient audio and speech can be compared. The dominant approach trains one unified tower across all modalities. This works, but it couples every capability to every other: adding a modality, updating one encoder, or diagnosing a failure means retraining or re-measuring the whole model. Strong specialist encoders already exist for each modality. The question we ask is whether a shared space can be constructed from them rather than learned jointly, and what such a construction gains and loses.

There are two challenges in achieving this. Specialist encoders produce geometries that are not compatible, and choosing one of them as the shared space discards what the others capture: a text encoder gives strong text retrieval but no vision alignment, and a vision–language encoder gives the reverse. And practical retrieval systems truncate embeddings to save memory. If a second geometry is simply appended to the first, any truncation to half width or less removes it entirely (Figure 2a).

OMNIROUTER addresses both. A closed-form ridge mapping between the two text spaces defines a residual that holds what the vision-aligned encoder knows and the text encoder does not (§3.2). The two geometries are interleaved coordinate-wise so that every prefix retains both (§3.3). Each modality reaches the shared space through its own small projector, so the similarity score decomposes into per-block terms and a change to one route cannot move another modality's scores (Eq. 6).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

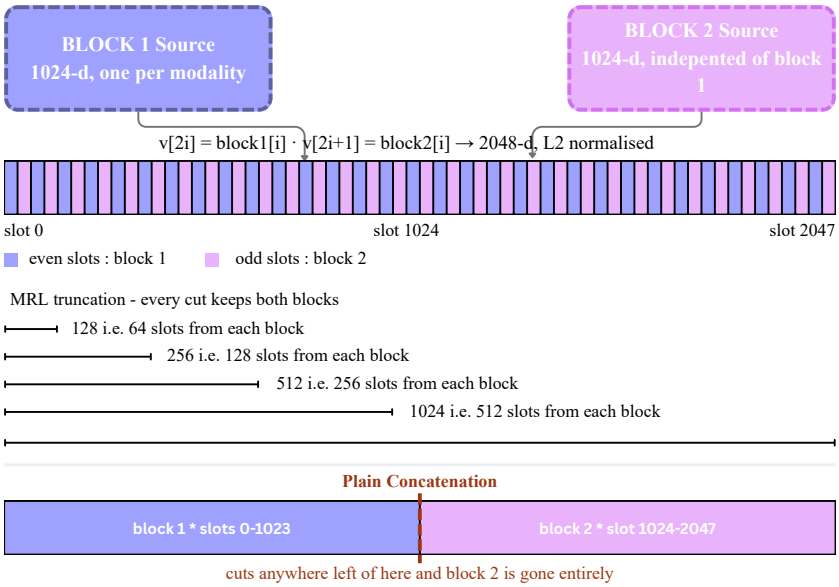

Figure 1: Illustrating interleaving two representation blocks ensures that both remain present under Matryoshka truncation, unlike plain concatenation.

The construction is competitive with models that learn the space jointly. On five core directions OMNIROUTER exceeds LCO-Embedding-Omni, an 8.9B omni-modal embedder with five times its parameters, on every one, and against the size-matched jina-v5-omni three of five differences are established with no losses (§5.2). On a nineteen-task suite re-scored for twenty-one systems under one protocol it attempts 17 of 19 columns and leads more than any other system (§5.1).

Explicit routes also make failures diagnosable and repairs local. The six scorable speech tasks scored a mean NDCG@10 of 0.0109, which reads as a capability limit; it was a missing dispatch rule, and adding it raises the mean to 0.7657 with nothing retrained (§5.3). Adding a modality is local in the same sense: training audio into a text-and-image system takes 150 seconds and leaves every existing embedding bit-identical, whereas adapting a jointly trained tower to the same data moves its text and image embeddings (§5.4).

Our contributions are:

- **A way to compose two independently trained embedding geometries without joint training.** A closed-form ridge map plus a residual gives a second block that holds what SigLIP knows and Qwen3 does not, and measure what each geometry contributes (§3.2, §5.5).
- **An interleaved coordinate layout** that keeps both geometries under Matryoshka truncation, with a measurement of where it helps and where it does not (§3.3, §5.5).
- **A modular encoder built from these two ideas**, in which routes are explicit and each modality's projector can be trained, replaced or added without changing any other modality's scores (§3.4, §5.4).

2 RELATED WORK

The space that multimodal retrieval needs has been built in two ways. The first is to learn it jointly: CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) train image and text encoders together contrastively, CLAP (Elizalde et al., 2023; Wu et al., 2023) does the same for audio, and text embedding models learn a shared geometry within one modality (Chen et al., 2024; Wang et al., 2022). Omni-modal encoders extend this to four or more modalities in one network, by pretraining a joint tower (Chen et al., 2023) or by contrastively refining a language-centric backbone so that a single decoder emits every modality's embedding (Xiao et al., 2025; Xu et al., 2025). These models are

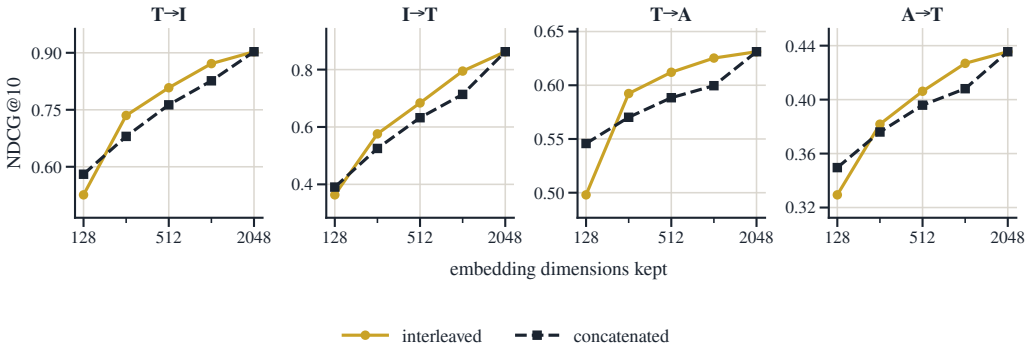

Figure 2: Measured NDCG@10 as the vector is truncated vs. concatenated, one panel per retrieval.

strong, but the space they learn is a property of the whole model: adding a modality, replacing an encoder or repairing a failure touches shared weights and forces everything else to be re-measured.

The second way keeps specialist encoders and connects them. ImageBind (Girdhar et al., 2023) anchors every modality to a frozen image space, and LanguageBind (Zhu et al., 2024), UniBind (Lyu et al., 2024) and OmniBind (Wang et al., 2025) anchor to a language or hub space instead, with parameter-efficient adapters (Houlsby et al., 2019; Hu et al., 2022) making the connectors cheap. jina-embeddings-v5-omni (Hönicke et al., 2026) is closest to ours: it freezes the text model and the added encoders, trains only connectors, and so leaves its text embeddings unchanged when modalities are added. What these share is a *single hub geometry*, every specialist is projected into one space, and whatever the hub does not represent is lost. We keep two geometries and define the second as a residual of the first, so that both survive.

Defining that residual requires a map between two spaces that were never trained together. Cross-lingual work showed that a single linear transformation recovers much of the alignment between separately trained word embeddings (Mikolov et al., 2013; Xing et al., 2015; Smith et al., 2017; Lample et al., 2018), and model stitching (Lenc & Vedaldi, 2015; Bansal et al., 2021) and relative representations extend the idea to neural representations more broadly. We use the same assumption across two text encoders rather than two languages, and only to define what the second block should *remove*. Keeping two geometries then creates a storage problem. Retrieval indices are routinely truncated, and Matryoshka Representation Learning (Kusupati et al., 2022) makes this possible by training embeddings whose prefixes remain useful, but it assumes a single space whose coordinates are ordered by importance, and when two blocks are concatenated that assumption fails, because every prefix up to half width belongs to one block. We interleave and measure the effect in §5.5.

A constructed space still has to be reached, and speech is where the route matters most: spoken audio can be embedded acoustically or through a transcript (Radford et al., 2023), and the two answer different questions. A design that spans four input families in one index also needs an evaluation that spans them fairly, and this is where existing practice is weakest. Benchmarks such as MTEB (Muennighoff et al., 2023) and BEIR (Thakur et al., 2021) establish leak-free retrieval protocols, which we follow (Karpukhin et al., 2020; Wang et al., 2022), but systems are usually compared under their own evaluation code and per-family leaderboards, and a modality a model cannot encode is easily recorded as a zero. We re-score every system with one evaluator and keep a task a model cannot attempt distinct from one it attempts and fails (§5.1).

3 Methodology

3.1 Overview

Figure 3 shows the complete system, and four terms name its parts. A **tower** is one of the four frozen pretrained encoders: Qwen3-Embedding-0.6B for text (596M), SigLIP-large-patch16-384 (Zhai et al., 2023) for images and for a second text embedding (652M), FineLAP for ambient audio (244M), and Whisper-small (Radford et al., 2023) for speech (242M), which is used only to produce transcripts and never as a feature extractor. A **block** is one of the two 1024-dimensional halves of

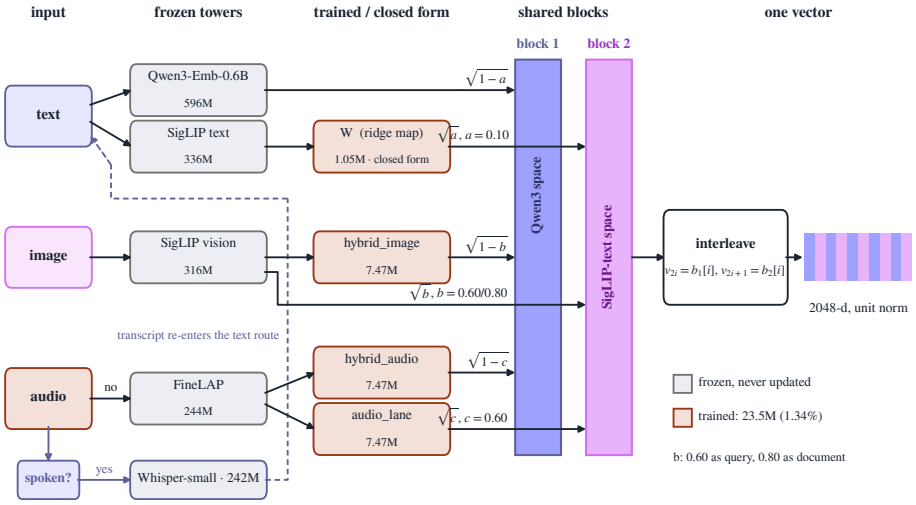

Figure 3: **Architecture and dispatch.** *Frozen* (grey): four pretrained towers. *Trained* (orange): image projector, two 7.47M audio projectors, and closed-form ridge map W. *Dispatch*: text/images use their own lanes; spoken audio is transcribed into the text route (dashed), requiring no trainable parameters or speech data, while other audio uses the FineLAP ambient lane.

the output vector; there are exactly two and every modality writes into both, block 1 being Qwen3 space and block 2 the SigLIP-text residual. A **lane** is the path one modality takes into the blocks, tower, projector and mixing weight, and a **route** is the fixed dispatch decision that sends an input to a lane. Freezing is not only an efficiency choice: it fixes each block's geometry, which is what makes the independence result of §3.3 possible. Only 23.5M parameters are trained, three projectors of 7.47M and a 1.05M closed-form map, 1.34% of the 1.76B resident model; a text query touches 0.93B of it and an image query 0.32B (Appendix A).

3.2 Two geometries

OmniRouter uses two representation spaces rather than one. A single space would mean choosing either geometry: the text space discards what SigLIP captures and Qwen3 does not, and the vision space sacrifices the text quality that carries T→T retrieval. Two blocks keep both, at cost of width.

Block 1 is Qwen3 space, used unchanged. **Block 2** is the *SigLIP-text residual*. Let $\mathbf{q}$ be a Qwen3 text embedding and $\mathbf{s}$ the SigLIP text embedding of the same string. With $\mathrm{nz}(x) = x/\|x\|_2$ and W mapping Qwen3 space to SigLIP space,

$$\ell(\alpha) \;=\; \mathrm{nz}(\mathbf{s} - \alpha\,\mathbf{q}W)\,. \tag{1}$$

The subtraction removes what Qwen3 already explains, leaving what SigLIP captures and Qwen3 does not; without it the blocks would be largely redundant. We compute W once in closed form,

$$W = (Q^\top Q + \lambda I)^{-1} Q^\top S, \qquad \lambda = 1.0, \tag{2}$$

where Q and S hold paired Qwen3 and SigLIP text embeddings from COCO (Lin et al., 2014). Closed form removes a training loop, optimizer and a seed from the reproduction path, and a linear map between independently trained spaces has precedent in cross-lingual alignment (Lample et al., 2018).

Each modality then produces two half-width blocks, combined with a scalar weight:

$$\text{text} = \left[\sqrt{1-a}\,\mathbf{q}\,;\ \sqrt{a}\,\ell(\alpha)\right], \tag{3}$$

$$\text{image} = \left[\sqrt{1-b}\,h_{\text{img}}\,;\ \sqrt{b}\,\mathbf{s}_{\text{img}}\right], \tag{4}$$

$$\text{audio} = \left[\sqrt{1-c}\,h_{\text{aud}}\,;\ \sqrt{c}\,\ell_{\text{aud}}\right], \tag{5}$$

where the query and document sides may use different weights (a_q, a_d and b_q, b_d). Here h_{img} and h_{aud} are trained projectors into Qwen3 space, $\mathbf{s}_{\text{img}}$ the native SigLIP image embedding and ℓ_{aud} a trained projector into the residual space. The square-root weighting preserves the norm, so a weight changes the balance between blocks and not the scale. We use $a_q = a_d = 0.10$, $b_q = 0.60$, $b_d = 0.80$, $c = 0.60$, $\alpha = 0.75$; the image sides are deliberately asymmetric, so a caller must declare which side it encodes. Every constant was selected on data no reported number uses, b_q on 3,000 held-out COCO rows, c on a held-out AudioCaps-train split, a and α on held-out SciFact and COCO, and every sweep has an interior optimum with both endpoints clearly worse (Appendix B).

3.3 Projectors and coordinate layout

Projectors. Each projector maps a pooled 1024-d tower feature to 1024-d and has 7.47M parameters. A pooled feature has no token sequence, so the projector reshapes it into 16 chunks of 64 and runs a self-attention branch (Vaswani et al., 2017) and a state-space branch over those chunks in parallel, mixed by a sigmoid gate, added to an identity-initialised linear residual. The output projection is zero-initialised and the residual identity-initialised, so an untrained projector is exactly the identity map rather than noise: training departs from the frozen model's behaviour only deliberately, which matters because that behaviour is already strong for images. Whether the chunked sequence earns its place is an empirical question, and §5.5 answers it by comparing against parameter-matched MLPs: it earns little, and the training objective matters more.

Interleaving. We interleave the two blocks rather than concatenate them: $v[2i] = b_1[i]$ and $v[2i+1] = b_2[i]$. Interleaving is a permutation, so full-width scores are unchanged; the difference appears under truncation. A 512-dimensional prefix of an interleaved vector contains 256 coordinates of each block, whereas the same prefix of a concatenated vector contains block 1 only and effectively reverts to the pre-composition representation (Figure 2a). Because the two blocks occupy disjoint coordinates, their inner product contains no cross term:

$$\langle u, v \rangle = \sqrt{(1 - w_q)(1 - w_d)}\, \langle u_1, v_1 \rangle + \sqrt{w_q w_d}\, \langle u_2, v_2 \rangle. \tag{6}$$

Each term depends on one block only, so a modality's projector enters only that modality's blocks: changing it can alter no direction in which that modality does not appear. This is the formal basis of the isolation result, which we verify rather than assert in §5.4.

3.4 Routing

Routing is deterministic rather than learned (Algorithm 1, Appendix C). Text and images take their lanes. Audio is asked one of two different questions, what a clip sounds like, or what it says, so it has two routes: ambient audio goes to FineLAP, while speech is transcribed and the transcript follows the text path, producing a representation directly comparable with text and image embeddings without a separate index or similarity function. Two properties of this rule matter for reading our results. The speech/ambient decision is supplied with the input, in our evaluation, by the task, and is not detected from the waveform. And transcription inherits the ASR model's language set, so a language guard declines speech in a language Whisper does not cover instead of transcribing it as English.

4 Implementation Details

Objective. The audio projectors are trained with bidirectional InfoNCE (van den Oord et al., 2018) averaged over Matryoshka prefixes $\{128, 256, 512, 1024\}$ (Kusupati et al., 2022) at temperature 0.05 (Appendix D). The image projector is trained with a pairwise sigmoid loss plus distillation of SigLIP's native image–text ranking, which §5.5 shows is the better objective for a block that sits beside a frozen SigLIP block. Optimisation uses AdamW (Loshchilov & Hutter, 2019), learning rate 10^{-3}, weight decay 10^{-2}, gradient clipping at 1.0 and early stopping; all training runs from cached frozen-tower features and fits on one 48 GB GPU.

Training data. The image lane is trained on 82,783 COCO training pairs (Lin et al., 2014) and the audio lane on 45,178 AudioCaps pairs (Kim et al., 2019), roughly 128k pairs in total. We use **no speech training data**: the speech result of §5.3 uses zero speech supervision, because speech is not a lane but a route into an existing one.

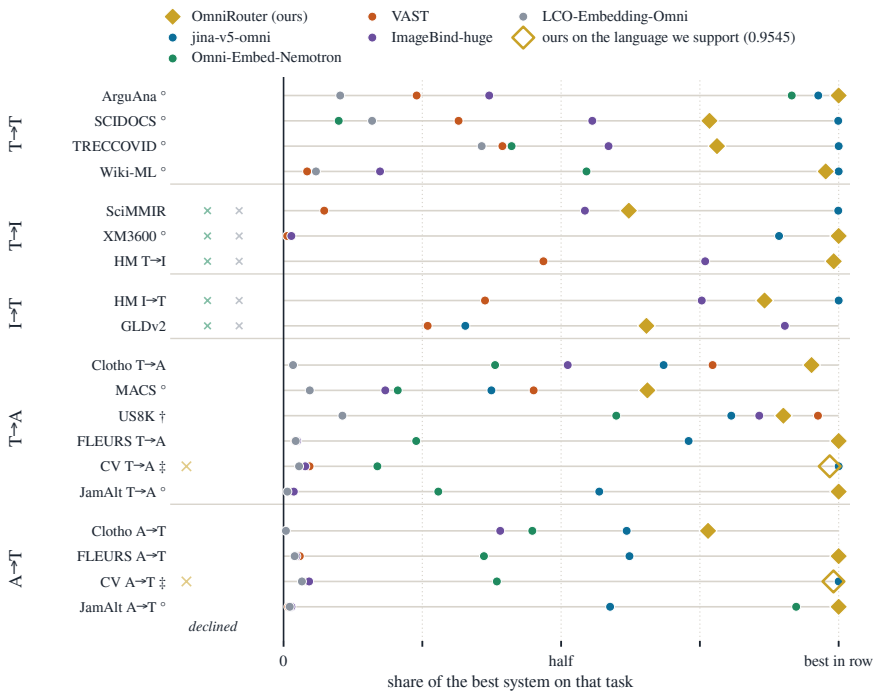

Figure 4: **Every suite task, six systems, one protocol.** Rows are grouped by retrieval direction and normalized by each task's best system, as task ceilings vary widely; e.g., UrbanSound8K has 10 queries/5,090 clips (ceiling $\approx 1/500$), which would be flattened on a shared absolute scale. Raw NDCG@10 appears in Tables 10a–10d. Unsupported tasks have no dot and are shown in the gutter left of zero. **Hollow diamonds are the sole exception:** for the Common Voice (‡) Abkhaz rows, our language guard declines, marked $\times$ in the gutter; the diamond instead shows our supported English score (0.9545 both directions), relative to jina's $0.9701/0.9635$ (Appendix I). Diamonds are therefore not comparable to the solid Abkhaz dots.

Protocol. Every query is embedded from its own modality alone and no corpus content passes through the query path; we report NDCG@10. Every baseline is called through the path its model card documents, including separate query and document entry points where a card defines them, omitting that understates such models by up to 0.06 NDCG@10 on a text task, which we corrected.

5 Results and Analysis

Every number reports one configuration: the deployed model, with the audio skip lane at its tuned weight $c = 0.60$, fixed on held-out data before the report set was scored.

5.1 Coverage and Comparison

Can a constructed space compete across every input family when all systems are measured the same way? Published numbers for omni-modal embedders come from different evaluation code, prompts and task subsets, so we re-score twenty-one systems on a nineteen-task suite with one evaluator: four text tasks, five image, four ambient-audio and six speech (Figure 4; Appendix J gives the raw scores by family).

Blanks are kept in three kinds and none is a zero: **n/a**, the system cannot encode that modality; **gap**, it has the modality but no route for this instance, our own two blanks, where the suite loads Common Voice's Abkhaz configuration and our guard declines rather than transcribing it as English. Scoring any of these as 0.0000 would let a system that never attempted a task read as one that tried and failed. Four encoders that are n/a on every column, none with a text side, are excluded.

CLAP leads ambient audio (0.2769 on Clotho against our 0.2634): our audio lane is a 7.47M projector on a frozen encoder, against a full audio–text model built for that task. jina takes the two text

Table 1: **Five retrieval directions and paired significance tests.** NDCG@10 across ArguAna, Flickr30k, and AudioCaps. † denotes our 1024-d variant; best scores are bold. The last row gives paired differences vs. jina-v5-omni (B=10,000); 95% CIs excluding zero indicate significance.

System	Params	Dim	T→T	T→I	I→T	T→A	A→T
OMNIROUTER	1.76B	2048	**0.5829**	0.9028	0.8623	**0.6312**	**0.4355**
OMNIROUTER†	1.76B	1024	0.5739	0.9008	0.8476	0.6125	0.4297
jina-v5-omni-small	1.74B	1024	0.5731	0.8914	0.8545	0.4607	0.3060
Omni-Embed-Nemotron	4.70B	2048	0.5320	0.6473	0.3319	0.1319	0.0887
VAST	1.40B	512	0.1416	**0.9089**	**0.8695**	0.4311	0.3469
ImageBind-huge	1.20B	1024	0.2157	0.8676	0.8603	0.2247	0.1413
LCO-Embedding-Omni-7B	8.93B	3584	0.3688	0.8682	0.7913	0.4467	0.3085
Δ vs jina, 95% CI			+0.0098 [−0.003, +0.023] tie	+0.0113 [+0.006, +0.017] win	+0.0067 [−0.003, +0.016] tie	+0.1705 [+0.160, +0.181] win	+0.1294 [+0.111, +0.149] win

columns where a dedicated retrieval encoder beats a frozen 0.6B text tower we never train, and both Common Voice columns, where our guard declines. Each system leads where it is specialised; ours is eligible for every family at once.

Established: the constructed space is eligible across all four families and competitive within each, and its losses trace to what it does not train.

5.2 MULTIDIRECTIONAL MARGINS

Does routing through frozen specialists give up retrieval quality to models that learn the space? Table 1 scores six systems on five core directions and tests every margin against the size-matched competitor in the same table. OMNIROUTER leads T→T and both audio directions; VAST leads both image directions by 0.0061 and 0.0072.

Image is where these systems agree: the five highest T→I scores span 0.041, because every one of them sits on a strong frozen vision tower. *Audio is where they diverge*: we reach 0.6312 against jina's 0.4607 on T→A and 0.4355 against VAST's 0.3469 on A→T. These audio directions are AudioCaps, scored through the ambient lane; the speech result of §5.3 is a separate measurement on separate corpora and the two should not be read together.

Two margins are under 0.008, so we test them with a paired bootstrap of per-query differences ($B = 10{,}000$) against jina-v5-omni. Three exclude zero, T→I ($+0.0113$), T→A ($+0.1705$), A→T ($+0.1294$), and two do not, which we report as ties; none is a significant loss. Two bounds belong with this. The column winners and the established differences are not the same three: T→T is a column lead and a tie, T→I a column loss to VAST and a significant win over jina. And our vector is twice jina's width; at matched 1024 dimensions our purpose-built variant leads four of five as point estimates and trails on I→T. VAST's image margins sit inside our own interval half-widths, so neither is established in either direction.

Established: construction costs nothing measurable where every system shares a frozen vision tower and gains substantially on audio, where they differ; the largest model in the comparison gains nothing from its size.

5.3 SPEECH ROUTING

When a routed system fails, can the failure be located and repaired without retraining? Six of the suite's eight speech tasks are scorable by every system compared here, and they scored a mean NDCG@10 of 0.0109 (Table 8a, Appendix I), which is chance; read as a model result it says the architecture cannot represent speech. The routes say otherwise: spoken audio was falling through to the ambient lane, whose encoder represents timbre and acoustic events. Two recordings of different sentences by the same speaker in the same room are acoustically similar and semantically unrelated, so the lane was answering a question the task had not asked.

We added one dispatch rule, transcribe, then re-enter the text path. Nothing was trained, no parameter was added and no speech supervision exists anywhere in this system. The mean over the six scorable tasks rose from 0.0109 to 0.7657, a gain of $+0.7548$, and every scorable task improved. The mean is over six tasks, not eight: the two Common Voice columns are declined by the language guard rather than scored.

Two checks keep this from overstating. FLEURS partly measures transcription accuracy, since its corpus is the reference transcript of the query audio; the load-bearing evidence is SpokenSQuAD, whose corpus is not a transcript of the queries and still rises from 0.0053 to 0.5611, and GigaSpeech,

Table 2: **Design choices:** Top: retrieval with each geometry, both, and without residual subtraction.

Geometry	T→T	T→I	I→T	T→A	A→T
both blocks (deployed)	0.5829	**0.9028**	**0.8623**	**0.6312**	0.4355
block 1 only (Qwen3 space)	**0.5891**	0.8263	0.7136	0.5996	0.4081
block 2 only (SigLIP-text residual)	0.2851	0.8223	0.6509	0.5922	0.3885
both, no residual (α=0)	0.5846	0.8923	0.8279	0.6176	**0.4362**

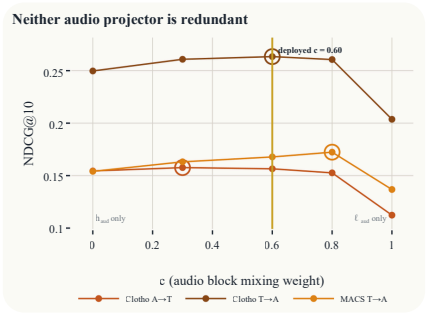

Figure 5: **Both audio projectors matter.** Sweeping $c \in [0, 1]$ on suite audio tasks. All tasks peak in $(0, 1)$; circles mark per-task maxima and the rule marks the deployed c=0.60, selected on held-out AudioCaps-train (Table 3).

which rises to 0.6221. And the route costs where its language is supported but competition is strong: on Common Voice's English configuration jina leads by 0.0156 and 0.0090 against our 0.9545 (Appendix I).

Established: in a unified model a low score is a verdict on the model; in a routed one it is first a question about which encoder received the input, and here the answer was a one-line repair worth +0.7548 mean NDCG@10.

5.4 Modality Isolation

Can a modality be added without disturbing the ones already deployed, and what does it cost? Equation 6 predicts that it can, and we test the prediction the way a deployment would meet it. Starting from a text-and-image system whose weight directory holds no audio checkpoint, we score the three directions that do not involve audio, train both audio projectors from scratch, add them to the same running system and score again (Table 6, Appendix G). Every text and image embedding is bit-identical afterwards, the largest change in any coordinate is exactly zero, so an existing index stays valid without re-encoding a single item, and the added modality works immediately, reaching 0.6323 and 0.4381. It cost 14.95M trained parameters and 150 seconds on one GPU.

For contrast we adapt a jointly trained tower to the same AudioCaps corpus by its own published recipe. The corpus holds 45,178 pairs; our own projectors hold out 3,000 of them for early stopping and train on 42,178, while the contrasts run one epoch over all 45,178, so if anything they see more of it than we do. VAST's single text encoder serves every cross-modal direction, so adapting it to audio moves the others: text-to-audio improves by +0.2211, confirming the adaptation took hold, while I→T falls by 0.0131 and T→I by 0.0200, and text and image embeddings drift by 0.22 (image) to 0.44 (audio) in mean cosine distance. LCO-Embedding-Omni-3B, whose LoRA adapts a decoder shared by every modality, moves by more: text-to-audio rises by +0.1729, T→I falls by 0.0353, and I→T and T→T move by +0.0135 and +0.0105. That two of those moves are improvements is not a defence of the design but the substance of the objection: a direction nobody retrained no longer returns the number that was measured for it, in either direction, and the published figure has to be earned again. Its text embeddings drift by 0.29 and its audio embeddings by 0.42 in mean cosine distance, against our largest coordinate change of exactly zero.

Established: a new lane is a local change in both the engineering and the measurement sense; where layers are shared across modalities it is not.

5.5 DESIGN CHOICES

Does each design choice earn its place, and where does it stop helping? Table 2 answers from a single cached encoding pass, so every difference comes from the variant alone; the deployed row reproduces an independent encoding run to four decimals.

Two geometries. On all four cross-modal directions both blocks beat the better single one, by 0.027 (A→T) to 0.149 (I→T), and removing the residual costs up to 0.034. On text-to-text the second geometry adds nothing, block 1 alone scores 0.5891 against 0.5829, which is the expected pattern: block 2 carries what a vision-aligned encoder knows and a text encoder does not, and a text-to-text query has no use for it.

Interleaving. Interleaving beats concatenation at 256, 512 and 1024 dimensions on all four cross-modal directions, by up to 0.081 (Figure 2b). It loses at 128, where each block keeps only 64 coordinates, and on text-to-text at every width, for the same reason as above: a concatenated prefix is pure Qwen3 space. The layout is the right default for a cross-modal index truncated to 256 dimensions or more, and the wrong one for a very small or text-only index.

Audio projectors. The two audio projectors are not redundant: their outputs agree at cosine 0.0165 on held-out AudioCaps, every task peaks strictly inside the c sweep, and the endpoints are asymmetric, removing the residual projector costs about 0.014 NDCG@10, removing the Qwen3-space projector about 0.06 (Figure 5; Appendix E).

Projector architecture and objective. Selected on held-out COCO with three seeds and the *composed* vector, the metric that matches deployment, since block 2 is frozen SigLIP, five architectures at a fixed objective span 0.009 on held-out T→I, and a parameter-matched MLP is within 0.004 of the Hybrid block. The objective matters more, and direction-dependently: distilling SigLIP's ranking wins I→T by 0.009 held-out and 0.052 on the report set, while InfoNCE wins held-out T→I by 0.011. A block-1-only metric ranks them the other way, which is why we do not use one (Appendix E).

Established: both geometries are needed for cross-modal retrieval and neither helps text-to-text; interleaving keeps them available below full width above 128 dimensions; and within this projector family the objective, not the architecture, is what matters.

6 LIMITATIONS

ASR coverage limits speech. ASR inherits its language set: Whisper covers 86/117 Common Voice languages, leaving 31 with no speech path. End-to-end audio degrades gracefully, while ours does not; larger ASR models raise the ceiling but not this structural limitation. Speech/ambient routing is also given as input, so deployment requires a classifier we do not provide. **Block 2 does not help text-to-text.** It is neutral to slightly harmful, and concatenation loses under truncation; text-only deployment is therefore better served by block 1. **Fine-tuning helps unevenly and costs vary by route.** Only Clotho is large, clean, and in-domain enough to show learning; HatefulMemes (676 training pairs) and GLDv2 (∼222 test items) mainly show no harm, while UrbanSound8K is degenerate. The image-lane null results from moving frozen tower hurts. Text is the most expensive route: block 2 adds SigLIP's text tower to Qwen3, so text touches 0.93B parameters, while speech additionally runs Whisper's decoder.

7 CONCLUSION

We asked whether multimodal retrieval requires joint learning. Our results show that key components can instead be constructed: a closed-form map between two text encoders defines a second geometry that improves every cross-modal direction, while coordinate permutation preserves it under truncation. The trade-offs are concrete: no graceful degradation beyond the ASR model's languages, no gain for text-only retrieval, and limited movement when frozen towers dominate the representation. More broadly, modularity makes route-specific behavior observable. In a unified tower, repairing one modality can affect others, as shown by image scores shifting after audio adaptation. With explicit routes, a new modality leaves existing indices intact. Future work should test whether multiple geometries can be added without linear growth in vector size, and whether speech/ambient routing can be learned without restoring the coupling that routing removes.

8 AI, Ethics and Reproducibility

AI use statement

In this work, we used generative AI tools for assistance with language editing, and improving the clarity and organization of the content. We also used generative AI tools for assistance with code for plotting and coding assistance. We did not use generative AI tools to generate experimental results. All AI-assisted content was reviewed by the authors.

Ethics statement

This work uses publicly available datasets and pretrained multimodal encoders for representation learning and retrieval evaluation. The study does not involve the collection of new human-subject data or the use of personally identifiable information. The proposed routing mechanism is evaluated for its ability to combine modality-specific representations while preserving the behavior of existing modality pathways. We are not aware of additional ethical concerns specific to the methods or experiments presented in this work.

Reproducibility statement

We provide detailed descriptions of the proposed routing architecture, frozen specialist encoders, training procedure, mapping and residual construction, embedding dimensions, and retrieval evaluation in the main paper and appendix. The appendix also reports the data splits, hyperparameter selection procedure, ablations, and computational requirements. Our experiments use standard publicly available retrieval benchmarks, including ArguAna, Flickr30k, and AudioCaps, together with publicly available pretrained encoders.

References

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 225–236, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/01ded4259d101feb739b06c399e9cd9c-Abstract.html.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2443–2463. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Advances in Neural Information Processing Systems*, volume 36, pp. 72842–72866, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/e6b2b48b5ed90d07c305932729927781-Abstract-Conference.html.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095889. URL https://doi.org/10.1109/ICASSP49357.2023.10095889.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023. doi: 10.1109/CVPR52729.2023.01457. URL https://doi.org/10.1109/CVPR52729.2023.01457.

Florian Hönicke, Michael Günther, Andreas Koukounas, Kalim Akram, Scott Martens, Saba Sturua, and Han Xiao. jina-embeddings-v5-omni: Geometry-preserving embeddings via locked aligned towers. *arXiv preprint arXiv:2605.08384*, 2026.

Neil Houlsby, Andrei Giurgiu, Stanisław Jastrzębski, Bruna Morrone, Quentin de Laroussilhe, Andréa Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011/.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham M. Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30233–30249, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c32319f4868da7613d78af9993100e42-Abstract-Conference.html.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H196sainb.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 991–999, 2015. doi: 10.1109/CVPR.2015.7298701. URL https://doi.org/10.1109/CVPR.2015.7298701.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer International Publishing, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Yuanhuiyi Lyu, Xu Zheng, Jiazhou Zhou, and Lin Wang. Unibind: Llm-augmented unified and balanced representation space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26752–26762, 2024. doi: 10.1109/CVPR52733.2024.02526. URL https://doi.org/10.1109/CVPR52733.2024.02526.

Tomas Mikolov, Quoc V. Le, and ılya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint*, 2013. arXiv:1309.4168v1.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2011–2037. Association for Computational Linguistics,

2023. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=r1Aab85gg.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=wCu6T5RIHgJ.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. URL https://arxiv.org/abs/2212.03533.

Zehan Wang, Ziang Zhang, Minjie Hong, Hang Zhang, Luping Liu, Rongjie Huang, Xize Cheng, Hengshuang Zhao, and Zhou Zhao. Omnibind: Large-scale omni multimodal representation via binding spaces. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l2izo0z7gu.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095969. URL https://doi.org/10.1109/ICASSP49357.2023.10095969.

Chenghao Xiao, Hou Pong Chan, Hao Zhang, Weiwen Xu, Mahani Aljunied, and Yu Rong. Scaling language-centric omnimodal representation learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=eA3Aum0rpA.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar (eds.), *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1006–1011, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1104. URL https://aclanthology.org/N15-1104/.

Mengyao Xu, Wenfei Zhou, Yauhen Babakhin, Gabriel Moreira, Ronay Ak, Radek Osmulski, Bo Liu, Even Oldridge, and Benedikt Schifferer. Omni-embed-nemotron: A unified multimodal retrieval model for text, image, audio, and video, 2025. URL https://arxiv.org/abs/2510.03458.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, and Zhifeng Li. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QmZKc7UZCy.

A Parameter accounting

Parameter counts are comparable across multimodal models only when they cover the same modalities, and VAST and jina do not include encoders outside the modalities evaluated here. We measure every count directly from the weights: VAST from its released checkpoint and ImageBind from its safetensors header. We report two counts because they measure different things. The *model total* includes every resident tower, 1.76B parameters, which is what a server accepting all four input types must hold. The *active per query* count includes only what encoding one query touches, so dispatch reduces it for some modalities: an image query uses 0.32B and an audio query 0.26B, since neither needs the text tower or the ASR model (Figure 6).

Two details matter. SigLIP ships image and text towers in one checkpoint and a query touches only one of them, so we count its 316.3M vision and 336.2M text parameters separately rather than charging the full 652.5M to both paths. And text queries are the expensive ones, not the cheap ones: block 2 is the SigLIP text embedding with the Qwen3 component removed, so a text query needs both text towers and touches 0.93B parameters, while a speech query adds the ASR model for 1.17B. Dispatch saves most on modalities with a single specialist and saves nothing on text.

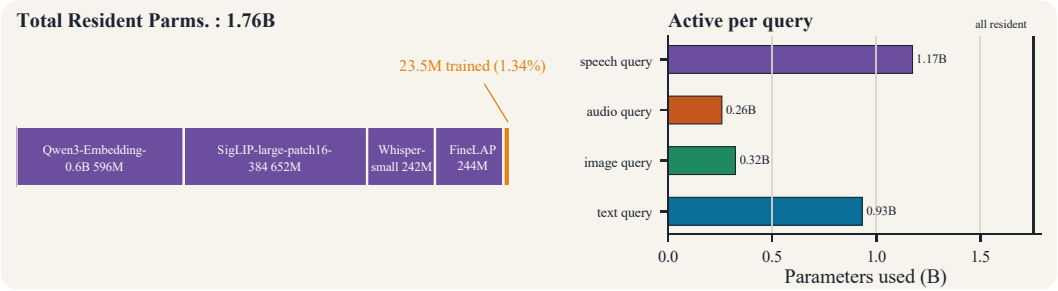

Figure 6: **Where the parameters are.** Left: the resident model by component. Right: what one query of each modality touches.

B Mixing weights were setting

Three scalars set how much of each block a vector carries: a for text, b for images and c for audio, with α fixing how much of the Qwen3 component is removed when block 2 is formed. None was chosen by hand or on the report set. Each was swept on held-out data, SciFact for a, a 3,000-row held-out split of COCO for b, and a held-out split of AudioCaps-train for c. Each was then frozen before any reported task was scored.

Table 3 gives the two sweeps that decide the cross-modal directions. Both optima are interior, which is the result that matters: at $b_q = 0$ an image vector is the projector alone and T→I falls to 0.6066; at $b_q = 1$ it is native SigLIP with no projector and falls further, to 0.4452; the best value is between

Table 3: **Mixing weights, set on held-out data.** Each weight was swept on a held-out split and fixed before the report set was scored; every optimum is interior, so neither block alone is best.

b_q	T→I	c	T→A	A→T
0.0	0.6066	0.00	0.3829	0.3319
0.2	0.6704	0.05	0.3946	0.3410
0.4	0.6913	0.10	0.3965	0.3430
0.6	0.6994	0.15	0.4003	0.3462
0.8	0.6915	0.20	0.4017	0.3474
1.0	0.4452	0.30	0.4043	0.3498
		0.40	0.4074	0.3518
		0.60	0.4116	0.3564
		0.70	0.4126	0.3564
		0.80	0.4132	0.3556
		0.90	0.4133	0.3521
		1.00	0.3638	0.3151

them. The audio sweep behaves the same way, rising from 0.3829 at $c = 0$ to a broad interior maximum and collapsing to 0.3638 at $c = 1$. Neither block alone is sufficient for a cross-modal direction, which is the claim Section 3.2 makes structurally and this table makes empirically.

We deploy $a = 0.10$, $\alpha = 0.75$, $b_q = 0.60$ with $b_d = 0.80$, and $c = 0.60$. The image weight splits by side because an image is encoded once as a query and once as a document and the two roles do not want the same mixture. The audio weight is deliberately not at the numerical peak of its sweep: $c = 0.60$ was fixed on the held-out split, and the sweep in Figure 5 is reported afterwards rather than used to choose. That distinction is the difference between a held-out value and a fitted one, and the caveat in Section 6 about c being tuned in-distribution for AudioCaps applies to it.

C DISPATCH RULE

Algorithm 1 is the whole routing rule, and it is deliberately short: every branch is a type test on the input, not a learned decision. Text and images take their own equations. Audio is the only input that branches twice, because *ambient* audio and *spoken* audio are different retrieval problems that happen to share a file format. An ambient clip is encoded acoustically, by what it sounds like. A spoken clip is transcribed and the transcript re-enters the text route, so what is retrieved on is what was said. Section 5.3 measures what that one branch changes.

The language guard on line 6 is the part that costs us coverage. The ASR model covers a fixed language set, and on an input outside it the system declines rather than transcribing as though it were English. A decline is not a failure and is not scored as zero anywhere in this paper: it appears as gap in Table 9 and as the two Common Voice rows excluded from the speech mean in Table 8. Table 8c scores the same two tasks on a configuration the guard accepts, so the cost of the guard can be read separately from the quality of the route.

Algorithm 1 Modality dispatch

1: **input** x, side $\in \{\text{query}, \text{doc}\}$
2: **if** x is text **then**
3: **return** Eq. 3 with (a, α) for that side
4: **else if** x is an image **then**
5: **return** Eq. 4 with b_q or b_d
6: **else if** x is *spoken* audio **then**
7: **if** language(x) unsupported **then return** decline
8: $t \leftarrow \text{Whisper}(x)$ {transcript, not acoustics}
9: **return** ENCODETEXT(t, side)
10: **else**
11: **return** Eq. 5 {ambient audio}
12: **end if**

Table 4: **Image projector architecture, selected on held-out data.** *Held-out* is T→I on the 3,000 held-out COCO rows, three seeds; *Composed* scores Flickr30k once inside the deployed vector, with InfoNCE rather than the deployed objective. The Hybrid block leads a parameter-matched MLP by 0.0026, and depth hurts.

Projector	Params	Held-out	Composed	
		T→I	T→I	I→T
Hybrid: attention ‖ SSM, gated (deployed)	7.47M	**0.6559**±0.0010	0.8906	0.8078
SSM branch only	2.12M	0.6536±0.0022	0.8885	0.8076
attention branch only	2.20M	0.6517±0.0011	0.8876	0.8035
MLP, 2 layers, parameter-matched	7.47M	0.6533±0.0014	0.8818	0.7771
MLP, 2 layers, 2048 hidden	5.25M	0.6510±0.0010	0.8812	0.7772
MLP, 4 layers, 2048 hidden	13.64M	0.6279±0.0020	0.8630	0.7296

Table 5: **Projector architecture and objective.** Selected on held-out COCO with three seeds, using the composed vector. The objective matters more than the architecture: a parameter-matched MLP is within 0.004 of the Hybrid block.

Projector	Objective	Params	Held-out, composed		Block 1	Report
			T→I	I→T	T→I	I→T
Hybrid	InfoNCE	7.47M	**0.7258**±0.0004	0.6905±0.0003	0.6560	0.8095
Hybrid	sigmoid + distillation	7.47M	0.7152±0.0005	0.6998±0.0008	0.6414	0.8619
SSM branch only	sigmoid + distillation	2.12M	0.7133±0.0014	**0.7027**±0.0005	0.6410	0.8668
attention branch only	sigmoid + distillation	2.20M	0.7116±0.0008	0.7004±0.0017	0.6356	0.8630
MLP, parameter-matched	sigmoid + distillation	7.47M	0.7112±0.0012	0.6979±0.0007	0.6380	0.8566
MLP, 4 layers	sigmoid + distillation	13.64M	0.7062±0.0009	0.6873±0.0024	0.6175	0.8437

D TRAINING OBJECTIVE

The loss is applied at several prefix lengths rather than only at full width. That is what makes the deployed vector truncatable: a prefix is trained to be a usable embedding in its own right, not merely the first half of one. Section 3.3 explains why the prefix must interleave the two blocks for this to mean anything, and Figure 2b measures what a prefix retains.

Only the audio projectors use this objective. The deployed image projector is trained with a sigmoid loss plus rank distillation from SigLIP instead, which is the configuration selected on held-out data in Table 5; the two objectives disagree by direction, and the appendix section below reports that disagreement rather than hiding it. Training runs from cached frozen-tower features, so no tower is loaded during projector training and a full run fits on one GPU.

The audio projectors minimise the Matryoshka InfoNCE loss

$$\mathcal{L} = \frac{1}{|D|} \sum_{d \in D} \frac{1}{2} \left[\mathrm{CE}\left(\frac{\mathbf{t}^{(d)} \mathbf{p}^{(d)\top}}{\tau}, I_N \right) + \mathrm{CE}\left(\frac{\mathbf{p}^{(d)} \mathbf{t}^{(d)\top}}{\tau}, I_N \right) \right] \tag{7}$$

where $\mathbf{x}^{(d)}$ is the first d dimensions of $\mathbf{x}$, $\mathbf{t}$ the text target, $\mathbf{p}$ the projected modality and $\tau = 0.05$.

E PROJECTOR ARCHITECTURE AND OBJECTIVE

Two tables settle the image projector. Table 4 holds the objective fixed and varies the architecture; Table 5 then varies the objective. Both are selection experiments, not leaderboards. In each, every row trains the image projector on COCO with the 3,000 held-out rows removed, three seeds per row, and is ranked by retrieval with the *composed* vector at deployed weights on those held-out rows. The composed metric is the one that matches deployment, because block 2 of an image vector is frozen SigLIP no matter what the projector learns; scoring the projector in isolation measures something the system never uses.

Table 6: **Adding a modality: what moves, and its costs.** *Top:* audio added to a text-and-image system; max $|\Delta|$ is the largest change in any coordinate of any existing embedding. *Bottom:* two unified models adapted to the same audio by their own recipes, one epoch, so their drift is a lower bound.

	T→T	T→I	I→T	T→A	A→T
OMNIROUTER*: add audio to a text+image system*					
before adding audio	0.5834	0.9028	0.8622	n/a	n/a
after adding audio	0.5834	0.9028	0.8622	0.6323	0.4381
max $\|\Delta\|$, any embedding	0	0	0		
VAST, adapted to audio					
Δ NDCG@10	+0.0629	-0.0200	-0.0131	+0.2211	
LCO-Embedding-Omni-3B, adapted to audio					
Δ NDCG@10	+0.0105	-0.0353	+0.0135	+0.1729	

Cost of adding audio to OMNIROUTER: 14.95M trained parameters and 150 s of training on one GPU, from cached tower features; no frozen parameter updated and no text or image item re-encoded.

That distinction changed the outcome, which is why the table keeps both columns. Ranked by a block-1-only metric, InfoNCE comes first. Ranked by the composed vector, and independently on the report set, the sigmoid-plus-distillation objective wins I→T. We report the selection metric we used and the one that disagrees with it, because the disagreement is the finding: within this family of projectors the objective decides more than the architecture, and a parameter-matched MLP lands within 0.004 of the Hybrid block while five architectures at a fixed objective span 0.009 on held-out T→I. Depth does not help.

F AUDIO PROJECTORS

The audio route carries two projectors, one into each block, and the c sweep tests whether the second one does anything. Its endpoints are the two ablations: $c = 0$ keeps only the Qwen3-space projector and $c = 1$ keeps only the residual-space one. Both endpoints fall below the deployed $c = 0.60$ in Table 3, so neither projector is redundant, and they fall by different amounts. Dropping the Qwen3-space projector ($c = 1$) costs more than dropping the residual one ($c = 0$), and Figure 5 shows the same interior peak on every suite audio task. Their outputs agree at cosine 0.0165 on held-out AudioCaps, which is to say they have learned nearly orthogonal things about the same clip.

G ADDING A MODALITY

Table 6 reports the add-a-modality test in three stages. We start from a system whose weight directory contains no audio checkpoint at all, score the three directions that do not involve audio, train both audio projectors from scratch, add them to the same running system, and score again. The top block is that before-and-after pair.

Max $|\Delta|$ is the number the claim rests on: the largest change in any coordinate of any query or document embedding across the whole evaluation. It is exactly zero, not approximately zero, which follows from Equation 6: a route that was not touched contributes the same per-block terms it did before. An existing index therefore stays valid without re-encoding a single item. The added modality works immediately, and the cost is stated beneath the table.

The bottom block is the contrast that makes the result meaningful. Two jointly trained models are adapted to the same audio pairs by their own published recipes, and because one encoder serves every direction in those models, adapting to audio moves text and image embeddings as well. Both contrasts ran a single epoch rather than the full recipe, so their drift is a lower bound on what a full adaptation would move, not an exaggeration of it.

Table 7: **Fine-tuning one lane at a time.** Selected on held-out validation and scored once. Δ is against the frozen model; OOD is a corpus never trained on.

Corpus	Lane	Pairs	In-domain Δ	OOD Δ	Final
Clotho	audio	3,835	+0.0557 / +0.0604	+0.0044	learns
HatefulMemes	image	676	+0.0005 / +0.0057	-0.0015	neutral
GLDv2	image	1,545	-0.0127	-0.0013	no gain
SciMMIR	image	5,000	-0.0055	-0.0087	harmful
UrbanSound8K†	audio	4,053	+0.0009 / -0.0084	-0.0095	excluded

H Fine-tuning the trainable fraction

Table 7 asks how far the trainable 1.34% can be pushed on a new corpus. One lane is fine-tuned at a time, warm-started from the deployed checkpoint, selected on held-out validation and scored once on test. Δ is against the frozen model, and the OOD column is a corpus the lane never saw, so a lane that improves in-domain while destroying the OOD column has overfitted, not adapted.

The audio lane adapts and the image lane does not, on any corpus we tried. We proposed an explanation: block 2 of an image vector is native SigLIP at weight 0.60–0.80, so most of the embedding is a frozen tower the projector cannot move. We then tested that by setting $b = 0$, which hands the projector the entire vector. Fine-tuning is still worse at $b = 0$. The explanation therefore fails its own test and we withdraw it, leaving the asymmetry recorded and unexplained rather than furnished with a story that does not survive a check.

I Speech: per-task results and a supported-language comparison

Table 8 gives the speech results at full resolution, in three parts. Subtable (a) puts every speech task under the same frozen weights through both routes: the acoustic lane, whose encoder represents how a clip sounds, and the ASR route of Algorithm 1. Every task that can be scored improves, and the mean is taken over the six scorable tasks, not over all eight. The two Common Voice rows are declined by the language guard and are excluded rather than counted as zero, which would flatter the gain.

Subtable (b) places the routed system against baselines on the same tasks, so the improvement is not read only against ourselves. Subtable (c) is the counterweight to the headline: on the Common Voice configuration whose language the ASR model does support, jina leads us on both directions. The route is what carries the gain in Subtable (a), and where a dedicated system competes on its own terms it does not win by default.

J Every system, every suite task

This section holds the complete suite at one protocol. Table 9 lists every system, what it attempted per input family, and how many columns it led; Table 10 gives the underlying scores split by family; and Figure 4 in Section 5.1 shows the same numbers as one picture.

Three conventions matter when reading these. First, a blank is never a zero, and the three kinds are kept apart: **n/a** is a modality a system cannot encode, gap is a modality it has but no route for this instance, and **err** is an attempt that failed. Scoring any of them as 0.0000 would let a system that never attempted a task read as one that tried and failed. Second, every system is called through the query and document entry points its own model card documents, not through a single generic call, because several of these models score substantially differently on the two sides. Third, we do not average a system over the subset of columns it happens to cover; breadth and score are reported separately, since a mean over a convenient subset rewards narrowness.

Table 8: **Speech, per task.** Every speech task under the same frozen weights: the acoustic lane, the ASR route, and the one configuration where the route's language is supported and the comparison is like for like.

(a) Acoustic lane against the ASR route.

Task	Acoustic	ASR dispatch	Δ
FleursA2T	0.0160	0.9643	+0.9483
FleursT2A	0.0219	0.9481	+0.9262
JamAltLyricA2T	0.0058	0.7532	+0.7474
JamAltLyricT2A	0.0150	0.7456	+0.7306
GigaSpeechT2A	0.0017	0.6221	+0.6204
SpokenSQuADT2A	0.0053	0.5611	+0.5558
CommonVoiceT2A	0.0181	gap	–
CommonVoiceA2T	0.0232	gap	–
mean	**0.0109**	**0.7657**	+0.7548

(b) Against baselines on the same tasks.

Task	Ours	jina-v5-omni	Δ
FleursA2T	0.9643	0.6009	+0.3634
JamAltLyricT2A	0.7483	0.4240	+0.3243
JamAltLyricA2T	0.7532	0.4431	+0.3101
FleursT2A	0.9481	0.6918	+0.2563
SpokenSQuADT2A	0.5611	0.4270	+0.1341
MACST2A	0.1678	0.0958	+0.0719
CommonVoiceA2T‡	0.9545	0.9635	-0.0090
GigaSpeechT2A	0.6221	0.7854	-0.1633
CommonVoiceT2A‡	0.2363	0.9701	-0.7338

(c) Common Voice, English configuration.

Direction	Ours (ASR)	jina-v5-omni	Δ
T→A	0.9545	0.9701	-0.0156
A→T	0.9545	0.9635	-0.0090

Table 9: **Every system's suite coverage.** Columns attempted per input family, total, and columns led. A family a system cannot encode is an absence, never a zero.

System	Params	Txt	Img	Amb	Spch	Total	Leads
OMNIROUTER (ours)	1.76B	4/4	5/5	4/4	4/6	17/19	**6**
jina-v5-omni-small	1.74B	4/4	5/5	4/4	6/6	19/19	**5**
jina-v5-omni-small-retrieval	–	4/4	5/5	4/4	6/6	19/19	**2**
jina-v5-omni-nano-retrieval	–	4/4	5/5	4/4	6/6	19/19	0
nvidia-omni-embed-nemotron-3b	4.70B	4/4	5/5	4/4	6/6	19/19	0
ImageBind-huge	1.20B	4/4	5/5	4/4	6/6	19/19	0
LCO-Omni-3B	–	4/4	5/5	4/4	6/6	19/19	0
LCO-Omni-7B	8.93B	4/4	5/5	4/4	6/6	19/19	0
VAST	1.40B	4/4	5/5	4/4	6/6	19/19	0
CLAP-htsat-unfused	–	4/4	n/a	4/4	6/6	14/19	**3**
CLAP-larger-music-speech	–	4/4	n/a	4/4	6/6	14/19	**1**
CLAP-htsat-fused	–	4/4	n/a	4/4	6/6	14/19	0
CLIPA-ViT-H-14	–	4/4	5/5	n/a	n/a	9/19	**2**
openclip-ViT-H-14	–	4/4	5/5	n/a	n/a	9/19	0
openclip-ViT-L-14-DataComp	–	4/4	5/5	n/a	n/a	9/19	0
clip-vit-large-patch14-336	–	4/4	5/5	n/a	n/a	9/19	0
clip-vit-large-patch14	–	4/4	5/5	n/a	n/a	9/19	0
SigLIP-base-patch16-224	–	4/4	5/5	n/a	n/a	9/19	0
SigLIP-large	–	4/4	5/5	n/a	n/a	9/19	0
EVA-CLIP-8B	–	4/4	5/5	n/a	n/a	9/19	0
SigLIP2-large-patch16-512	–	4/4	5/5	n/a	n/a	9/19	0

Table 10: **The suite by input family.** NDCG@10 under one protocol, representative systems.

(a) Text-to-text.

System	ArguAna^o	SCIDOCS^o	TRECCOVID^o	Wiki-ML^o
OMNIROUTER (ours)	**0.5836**	0.1071	0.6924	0.8667
jina-v5-omni-small	0.5621	**0.1395**	**0.8867**	**0.8874**
nvidia-omni-embed-nemotron-3b	0.5342	0.0139	0.3642	0.4842
ImageBind-huge	0.2162	0.0776	0.5191	0.1542
LCO-Omni-7B	0.0595	0.0223	0.3165	0.0518
VAST	0.1399	0.0440	0.3496	0.0377
CLAP-htsat-unfused	0.0812	0.0078	0.1182	0.0423
CLIPA-ViT-H-14	0.1725	0.0808	0.6172	0.0503
SigLIP-large	0.1877	0.0034	0.0050	0.1810

(b) Image. XM3600 is multilingual, where English-only text towers collapse.

System	SciMMIR	XM3600^o	HM T→I	HM I→T	GLDv2
OMNIROUTER (ours)	0.2766	**0.5332**	0.7739	0.6692	0.5393
jina-v5-omni-small	**0.4443**	0.4760	0.7759	**0.7723**	0.2701
nvidia-omni-embed-nemotron-3b	0.2936	0.0819	0.1458	0.5102	0.3659
ImageBind-huge	0.2414	0.0075	0.5931	0.5819	0.7451
LCO-Omni-7B	0.0611	0.1677	0.0997	0.0871	0.3867
VAST	0.0326	0.0035	0.3656	0.2803	0.2141
CLIPA-ViT-H-14	0.2150	0.0046	**0.7808**	0.7647	**0.8250**
SigLIP-large	0.3181	0.5108	0.2233	0.0862	0.0413

(c) Ambient audio. UrbanSound8K gives 10 queries for 5,090 clips, so its ceiling is about $1/500$.

System	Clotho T→A	Clotho A→T	MACS^o	US8K†
OMNIROUTER (ours)	0.2634	0.1565	0.1678	0.0080
jina-v5-omni-small	0.1896	0.1265	0.0958	0.0072
nvidia-omni-embed-nemotron-3b	0.1055	0.0917	0.0527	0.0053
ImageBind-huge	0.1418	0.0799	0.0469	0.0076
LCO-Omni-7B	0.1063	0.0800	0.1845	0.0080
VAST	0.2140	0.1561	0.1153	0.0086
CLAP-htsat-unfused	**0.2769**	**0.2047**	**0.2217**	**0.0089**

(d) Speech, with our route active. gap is the language guard declining, not a failure.

System	FLEURS T→A	FLEURS A→T	CV T→A‡	CV A→T‡	JamAlt T→A^o	JamAlt A→T^o
OMNIROUTER (ours)	**0.9481**	**0.9643**	gap	gap	**0.7456**	**0.7532**
jina-v5-omni-small	0.6918	0.6009	**0.6545**	**0.7435**	0.4240	0.4431
nvidia-omni-embed-nemotron-3b	0.2265	0.3481	0.1106	0.2857	0.2078	0.6956
ImageBind-huge	0.0232	0.0221	0.0256	0.0342	0.0138	0.0107
LCO-Omni-7B	0.2854	0.2706	0.0695	0.0951	0.3948	0.2951
VAST	0.0198	0.0282	0.0306	0.0326	0.0117	0.0065
CLAP-htsat-unfused	0.0286	0.0211	0.0262	0.0126	0.0139	0.0060