Syn-Omni: Structured Specialization and Progressive Collaboration for Omnimodal Embeddings

Anonymous ACL submission

Abstract

Omnimodal embeddings naturally involve both shared representations and modality-specific features across heterogeneous inputs. However, existing omnimodal embedding methods often rely on a single shared parameter space over mixed-modality data, limiting structural separation between universal and modality-specific representations. To address this, we propose `Syn-Omni`, a unified framework for structured omnimodal adaptation with modality specialization and controlled cross-modal collaboration. Specifically, we introduce Orthogonal Modality-Expert LoRA (OME-LoRA), which decomposes adaptation into a shared LoRA path for universal semantics and modality-expert LoRA paths for modality-aware specialization. Furthermore, Progressive Synergy Routing (PSR) enables experts to first establish modality-specific priors, then gradually interact with other modality experts for cross-modal synergy. Evaluated across 81 diverse tasks spanning image, video, audio, and audio-visual modalities, `Syn-Omni` consistently outperforms omnimodal baselines, demonstrating the effectiveness of structured specialization and cross-modal progressive collaboration.

1 Introduction

Omnimodal Large Language Models (Omni-MLLMs) have opened a new frontier toward unified understanding across modalities, including text, images, video, audio, and their composed inputs (Xu et al., 2025a). With these versatile foundations, recent studies (Xiao et al., 2025; Chen et al., 2026) have successfully repurposed Omni-MLLMs into discriminative embedding models through contrastive fine-tuning (Jiang et al., 2025; Meng et al., 2026) with LoRA (Hu et al., 2022), offering an efficient and scalable framework for unified embedding spaces across heterogeneous modalities.

Despite their strong performance, existing omnimodal embedding methods largely rely on fully shared parameter adaptation over mixed-modality data. While such shared adaptation is effective in learning a unified representation space, it provides limited structural separation between modality-shared and modality-specific representations. Since different modalities exhibit both common semantic structures and distinct modality characteristics, a shared adaptation space may underutilize modality-aware specialization over the course of training.

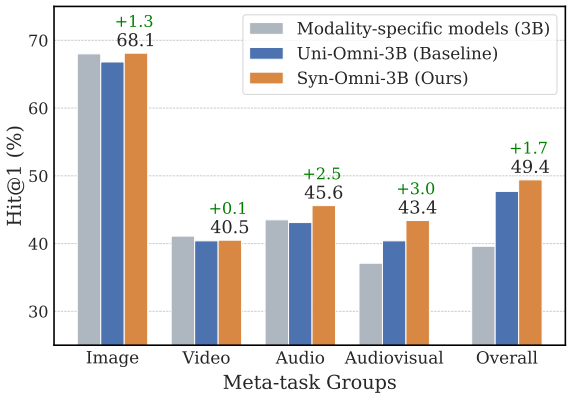

Figure 1: **Omnimodal adaptation strategies.** We compare three contrastive embedding setups: **(1)** modality-specific models, trained independently on each modality using a shared LoRA, **(2)** `Uni-Omni` (baseline), trained on mixed-modality data using a single shared LoRA, and **(3)** `Syn-Omni` (Ours), which combines a shared adaptation path with modality-specialized expert paths. While mixed-modality training enables strong unified representations, `Syn-Omni` further improves `Uni-Omni` baseline across diverse modality scenarios. We further expand the experimental results in Sec. C.2.

As shown in Fig. 1, different omnimodal adaptation strategies exhibit complementary strengths. Modality-specific models effectively preserve modality-aware characteristics, but their isolated training restricts cross-modal interaction and limits generalization. In contrast, `Uni-Omni`, a baseline trained on mixed-modality data with a single shared LoRA, learns a unified embedding space that performs well on composed multimodal set-

tings such as audiovisual tasks. However, such fully shared adaptation provides limited structural separation to model modality-specific characteristics, whereas completely isolated experts fail to leverage cross-modal synergy. These observations highlight the need for a structured omnimodal adaptation strategy that can jointly support a unified semantic foundation and modality-aware specialization while encouraging collaborative interactions within an omnimodal embedding space.

To this end, we propose `Syn-Omni`, a framework that structurally separates shared and modality-specialized expert paths while progressively enabling cross-modal collaboration. Our framework consists of two key components: *Orthogonal Modality-Expert LoRA* (OME-LoRA) and *Progressive Synergy Routing* (PSR). OME-LoRA separates adaptation into shared and modality-specific LoRA paths to capture both universal semantics and distinct characteristics. Complementing this, PSR gradually enables these isolated modality experts to interact and establish cross-modal synergy over the course of training. As shown in Fig. 1, `Syn-Omni` consistently improves upon the unified baseline, showing clear gains on composed multimodal tasks such as audiovisual understanding.

We extensively evaluate `Syn-Omni` across 81 diverse embedding tasks spanning image, video, audio, and audiovisual modalities, covering a broad range of task types including classification, retrieval, and QA. We show that `Syn-Omni` outperforms existing omnimodal embedding baselines while achieving more balanced performance across these diverse modality settings. These results demonstrate the effectiveness of structured specialization and progressive cross-modal collaboration for unified omnimodal representation learning.

We summarize our contributions as follows:

- We address the structural limitations of existing fully shared or entirely isolated omnimodal adaptation methods by presenting a unified framework that simultaneously balances robust semantic foundations and modality-aware specialization.
- We propose `Syn-Omni`, which combines OME-LoRA for structured decomposition of modality-shared and modality-specific adaptation with PSR for gradually enabling cross-modal collaboration across experts over the course of training.
- We evaluate `Syn-Omni` across 81 tasks spanning image, video, audio, and audiovisual modalities, covering diverse task types. Experimental results show that `Syn-Omni` consistently improves unified omnimodal baselines across diverse modalities.

2 Related Work

Omnimodal Embeddings. Multimodal Large Language Models (MLLMs) have facilitated the adaptation of these backbones into discriminative embedding models through contrastive adaptation. Early frameworks on vision-language models like E5-V (Jiang et al., 2024) and VLM2Vec (Jiang et al., 2025) pioneered this transition by applying LoRA (Hu et al., 2022) to text-image retrieval. Covering broader modalities, GVE (Guo et al., 2025) and VLM2Vec-V2 (Meng et al., 2026) extended this paradigm to video, while AuroLA (Xu et al., 2026) focused on audio-text retrieval. Furthermore, Omni-Embed-Nemotron (Xu et al., 2025b) and WAVE (Tang et al., 2026) enable any-to-any retrieval across text, image, audio, and video by leveraging large omnimodal backbones (Xu et al., 2025a). To further refine cross-modal alignment, methods such as LCO-Emb (Xiao et al., 2025) and e5-omni (Chen et al., 2026) introduce explicit calibration strategies, including language-centric refinement and modality-aware temperature scaling. While most existing approaches rely on fully-shared adaptation (*e.g.*, a single LoRA layer) across modalities, we introduce a decoupled adaptation framework with shared and modality-specific pathways while enabling cross-modal collaboration.

Structured Adaptation. Existing structured adaptation approaches suggest that explicitly separating parameters can improve specialization across domains. For instance, LIMoE (Mustafa et al., 2022) and MoDULA (Ma et al., 2024b) demonstrate that decoupling modality- or task-specific components leads to more effective learning. Beyond architectural separation, recent studies highlight limitations of naive cross-modal fusion and alignment strategies. DGL (Wei et al., 2025) shows that overly aggressive fusion can suppress modality-specific learning, while MCA (Wu et al., 2025) reports shortcut behaviors induced by simple contrastive alignment objectives. From an optimization perspective, Uni-X (Hao et al., 2026) further reveals the limitations of fully shared transformers in modeling heterogeneous modalities. Upon these findings, we design a structured adaptation framework that separates shared and modality-specific pathways while enabling controlled cross-modal interaction across experts via progressive routing.

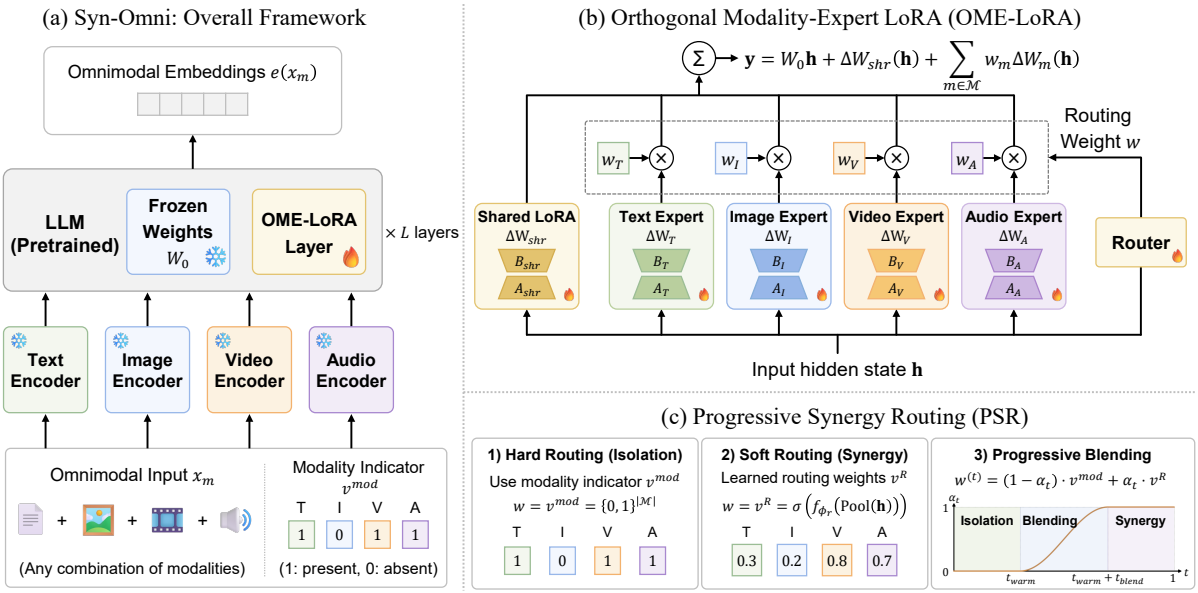

Figure 2: **Illustration of the `Syn-Omni` framework. (a)** Overall Framework: Omnimodal inputs x_m are encoded by modality-specific encoders and a pretrained LLM backbone with OME-LoRA to produce unified embeddings $e(x_m)$. **(b)** Orthogonal Modality-Expert LoRA (OME-LoRA): Shared LoRA path captures universal semantics, while modality-specific expert LoRAs model modality-aware characteristics. The outputs of modality experts are combined by the routing weight w. **(c)** Progressive Synergy Routing (PSR): PSR gradually shifts from initial modality-separate routing to later-stage collaborative routing, enabling progressive cross-modal synergy.

3 Syn-Omni Framework

We propose `Syn-Omni`, a unified framework that introduces structured adaptation beyond fully shared parameter space while enabling progressive cross-modal synergy across experts. As shown in Fig. 2, our approach consists of two main components: *Orthogonal Modality-Expert LoRA* (OME-LoRA) and *Progressive Synergy Routing* (PSR). We first introduce our omnimodal embedding formulation `Uni-Omni` in Sec. 3.1. We then detail the design of OME-LoRA in Sec. 3.2, the PSR mechanism in Sec. 3.3, and the training objectives in Sec. 3.4.

3.1 Omnimodal Embedding Learning

We consider a mixed omnimodal training setting, where we construct a heterogeneous training mixture encompassing various modality combinations. Under this setup, we train `Uni-Omni`, a simple contrastive baseline that adapts a pretrained Omni-MLLM using a single modality-shared LoRA.

Omnimodal Representation. We define a base modality set $\mathcal{M} = \{T, I, V, A\}$, representing text, image, video, and audio, respectively. An input x_m is any non-empty subset $m \subseteq \mathcal{M}$, and we define a modality indicator $v^{mod} \in \{0,1\}^{|\mathcal{M}|}$ indicating the presence of each modality in x_m. To obtain the omnimodal embedding $\mathbf{e}(x_m)$ from a pretrained Omni-MLLM f_θ, the input is processed with an instruction prompt, and last-token pooling is applied from the final layer: $\mathbf{e}(x_m) = \mathrm{Pool}(f_\theta(x_m))$.

Contrastive Optimization via LoRA. The pretrained weights $W_0 \in \theta$ are frozen, while low-rank adapters ϕ are trained. For an intermediate hidden state $\mathbf{h}$, the adapted output is $\mathbf{y} = W_0\mathbf{h} + \Delta W_\phi(\mathbf{h})$. The model is optimized with an InfoNCE objective to align matched pairs in the unified embedding space. For a batch of N pairs, the loss is:

$$\mathcal{L}_{con} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\mathrm{sim}(\mathbf{e}_i, \mathbf{e}_i^+)/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(\mathbf{e}_i, \mathbf{e}_j)/\tau)}, \tag{1}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature hyperparameter. We further detail our omnimodal training mixture in Sec. A.1.

3.2 Orthogonal Modality Expert LoRA

The training of omnimodal embeddings involves heterogeneous modalities with different characteristics, yet typically relies on a single shared LoRA applied across all modalities. This motivates a more structured design to better organize shared and modality-specific adaptation. To this end, we propose *Orthogonal Modality-Expert LoRA* (OME-LoRA), which introduces a structured decomposition of shared and modality-expert LoRA paths.

Parallel Shared-Expert LoRA. Inspired by the decoupling of universal and domain-specific knowledge in multi-task learning literature (Ma et al., 2024b), we extend this foundation to the omni-modal setting to jointly model shared semantics and modality-specialized characteristics. Specifically, as shown in Fig. 2 (b), a shared LoRA path is consistently activated for all input modalities, while modality-expert LoRA paths are selectively engaged based on the modalities present in x_m. For an intermediate hidden state $\mathbf{h}$, the output of the OME-LoRA layer $\mathbf{y}$ is computed as follows:

$$\mathbf{y} = W_0\mathbf{h} + \Delta W_{shr}(\mathbf{h}) + \sum_{m\in\mathcal{M}} w_m \Delta W_m(\mathbf{h}), \quad (2)$$

where ΔW_{shr} captures modality-shared parameter updates and ΔW_m models modality-specific counterparts. The routing weights w_m enable adaptive integration of modality-expert outputs. We defer the details of modality-aware specialization in modality experts and the routing strategy for cross-modal synergy among experts to Sec. 3.3.

Decoupling via Orthogonal Penalty. To encourage separation between modality-shared and modality-specific paths, we introduce an orthogonal penalty $\mathcal{L}_{ortho}$ that enforces modality-specific experts to learn complementary weight directions relative to the shared pathway. Given that each LoRA update is decomposed into low-rank matrices $\Delta W = BA$, we define the vectorized form of the input-side projection matrix as $\mathbf{a} = \mathrm{vec}(A)$. The orthogonality penalty is then defined as:

$$\mathcal{L}_{ortho} = \frac{1}{|\mathcal{M}|} \sum_{m\in\mathcal{M}} \left(\frac{|\tilde{\mathbf{a}}_{shr} \cdot \mathbf{a}_m|}{\|\tilde{\mathbf{a}}_{shr}\| \|\mathbf{a}_m\|} \right)^2, \quad (3)$$

where $\mathcal{L}_{ortho}$ minimizes the cosine similarity between the weight directions of the shared and expert LoRA parameters. To stabilize this objective, we treat the shared component as a fixed reference by applying a stop-gradient operator, $\tilde{\mathbf{a}}_{shr} = \mathrm{sg}[\mathbf{a}_{shr}]$, where $\mathrm{sg}[\cdot]$ blocks gradient propagation. In this way, shared and modality-specific parameters are encouraged to occupy complementary directions, promoting structured specialization without affecting the shared representations.

3.3 Progressive Synergy Routing

The interaction between modality experts is governed by the routing weights w_m in Eq. (2). While hard routing based on the modality indicator v^{mod} activates only the corresponding experts, it restricts the cross-modal knowledge exchange across experts. In contrast, while soft routing with a learnable router enables synergy, it can blunt expert specialization if the experts lack sufficient modality-aware priors early on, as explored in Fig. 5 (left).

To facilitate stable and progressive collaboration among experts, we propose *Progressive Synergy Routing* (PSR), which gradually transitions from modality-isolated hard routing to adaptive soft routing, allowing experts to first establish modality priors and then progressively exchanging complementary information across modalities, as illustrated in Fig. 2 (c). In addition, the Synergy Margin Loss (SML) is designed to avoid excessive suppression of absent-modality experts, preserving expert interactions necessary for cross-modal synergy.

Progressive Blending Mechanism. Specifically, we first introduce a learnable router f_{ϕ_r} that generates soft routing weights $\mathbf{v}^R = \sigma(f_{\phi_r}(\mathrm{Pool}(\mathbf{h})))$ from the intermediate hidden state $\mathbf{h}$, where f_{ϕ_r} is a fully-connected layer with parameters ϕ_r and $\sigma(\cdot)$ denotes the sigmoid function. This soft router enables experts from non-input modalities to contribute dynamically to the final outputs. To resolve the trade-off between expert specialization and synergy, we propose to progressively blend the hard and soft routing weights over the training progress t to *smoothly* shift from isolation to interaction. The blended routing weight $w_m^{(t)}$ is computed as:

$$w_m^{(t)} = (1-\alpha_t)\cdot v_m^{mod} + \alpha_t \cdot v_m^R, \quad (4)$$

where $\alpha_t \in [0,1]$ is scheduled along with the training progress ratio t (*e.g.*, current step / total steps):

$$\alpha_t = \begin{cases} 0, & 0 \le t < t_{\text{warm}} \\ g(t), & t_{\text{warm}} \le t < t_{\text{warm}} + t_{\text{blend}} \\ 1, & t_{\text{warm}} + t_{\text{blend}} \le t \le 1 \end{cases} \quad (5)$$

Here, $g(t) = \frac{1}{2}[1-\cos(\pi \frac{t-t_{\text{warm}}}{t_{\text{blend}}})]$ denotes a cosine-based monotonic transition function from 0 to 1. The resulting schedule consists of three phases: *warmup* ($t < t_{\text{warm}}$) for strict isolation with expert specialization, *Blending* ($t_{\text{warm}} \le t < t_{\text{blend}}$) for gradual blending of two routing weights, and *synergy* ($t \ge t_{\text{warm}} + t_{\text{blend}}$) for full dynamic collaboration via soft routing weights $\mathbf{v}^R$. As a result, the proposed mechanism successfully harnesses both modality-aware specialization and progressive cross-modal collaboration. To empirically capture this transition, Fig. 6 visualizes the evolution of routing weights over the course of training.

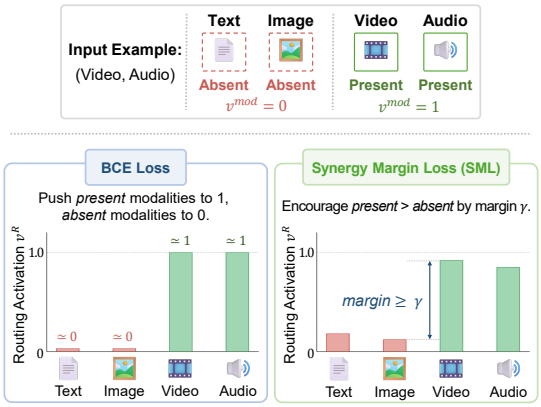

Figure 3: Conceptual comparison between BCE Loss and Synergy Margin Loss (SML). While BCE strictly suppresses absent modalities to 0, SML maintains a relative margin γ ($v^R_{\text{present}} - v^R_{\text{absent}} \geq \gamma$), which allows absent-modality experts to retain meaningful activations, preserving cross-modal synergy during training.

Synergy Margin Loss. To encourage modality-aware specialization in the router, we introduce an auxiliary supervision that guides the soft routing activations $\mathbf{v}^R$ using the modality indicators v^{mod}. While Binary Cross-Entropy (BCE) could directly enforce modality-aligned routing, it tends to excessively suppress absent modalities, eliminating potentially useful interactions across experts. To preserve such collaborative capacity while maintaining the dominance of input-relevant modalities, we propose a Synergy Margin Loss $\mathcal{L}_{SML}$. Let $\mathcal{P}$ and $\mathcal{A}$ denote the sets of present ($v_m^{mod} = 1$) and absent ($v_m^{mod} = 0$) modality indices in the input:

$$\mathcal{L}_{SML} = \frac{1}{|\mathcal{P}||\mathcal{A}|} \sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{A}} \max(0, v_j^R - v_i^R + \gamma), \tag{6}$$

where γ is a margin hyperparameter. As shown in Fig. 3, unlike BCE, the proposed objective enforces relative preference rather than absolute suppression, allowing related experts to retain meaningful activations while preserving specialization.

3.4 Training Strategy

We train `Syn-Omni` in an end-to-end manner by integrating OME-LoRA into the backbone and optimize the model with a joint objective that combines contrastive learning and structural regularization.

Architectural Integration. Instead of standard LoRA layers, we deploy OME-LoRA throughout the backbone, where each module consists of a shared adaptation path, modality experts, and a learnable router. Such learnable parameters are maintained independently at each layer.

Joint Optimization. The overall training objective $\mathcal{L}_{total}$ combines the primary contrastive objective with the proposed structural regularization losses:

$$\mathcal{L}_{total} = \mathcal{L}_{con} + \lambda_{ortho}\mathcal{L}_{ortho} + \lambda_{SML}\mathcal{L}_{SML}, \tag{7}$$

where λ_{ortho} and λ_{SML} control the strengths of the orthogonal penalty and Synergy Margin Loss, respectively. Crucially, both regularization terms $\mathcal{L}_{ortho}$ and $\mathcal{L}_{SML}$ are computed and averaged across all layers containing OME-LoRA modules.

4 Experiment

Training Data. We train our embedding models on a curated omnimodal data mixture of 1.12M samples spanning image, video, audio, and audiovisual modalities. The training mixture covers diverse tasks, including retrieval, classification, and QA, all reformulated into a unified contrastive learning framework. The training mixture covers various modality compositions, ranging from single-modality inputs to composed multimodal scenarios. Detailed dataset statistics and the complete training data composition are provided in Sec. A.1.

Model Variants. We build our `Syn-Omni` framework on top of pretrained Qwen-2.5-Omni (Xu et al., 2025a) and train both 3B and 7B scale variants. OME-LoRA modules are inserted into the LLM layers while keeping the remaining backbone parameters frozen. We additionally establish `Uni-Omni` as a controlled baseline trained with the same data and optimization settings, using only a single shared LoRA and contrastive objective.

Implementation Details. Images and videos are processed using the native resolution scaling strategy of Qwen-2.5-VL (Bai et al., 2025), while videos are uniformly sampled into 8 frames and audio signals are resampled to 16 kHz. Models are trained for 1 epoch using the AdamW (Loshchilov and Hutter, 2017) optimizer with a cosine learning rate schedule, a peak learning rate of 10^{-4}, a warmup ratio of 0.04, and a global batch size of 768. We employ GradCache (Gao et al., 2021) to maintain a large global batch efficiently. For the contrastive loss, we use a temperature τ of 0.02. Consistent with prior methods (Jiang et al., 2025), we pool on the EOS token position followed by L2 normalization to extract the final embeddings.

For OME-LoRA, we set the rank of the shared adaptation path to 8 and each modality expert rank to 4. We use a synergy margin γ of 0.2 and apply the proposed Progressive Synergy Routing

	Image (36)	Video (23)	Audio (12)	Audiovisual (10)	All (81)
3B Models					
Omni-Embed-Nemotron (Xu et al., 2025b)	44.1	36.5	24.5	28.5	33.4
LCO-Emb (Xiao et al., 2025)	58.1	**43.7**	42.1	35.5	44.8
e5-omni (Chen et al., 2026)	64.8	40.6	37.4	31.6	43.6
`Uni-Omni` (Contrastive baseline)	66.8	40.4	43.1	40.4	47.7
`Syn-Omni` (Ours)	**68.1**	40.5	**45.6**	**43.4**	**49.4**
7B Models					
OmniEmbed-v0.1 (Ma et al., 2025a)	44.2	35.0	31.6	29.0	35.0
OmniEmbed-Multivent (Zhan et al., 2025)	51.8	40.2	39.5	34.4	41.5
LCO-Emb (Xiao et al., 2025)	61.6	**45.5**	45.2	35.5	46.9
e5-omni (Chen et al., 2026)	**72.5**	44.3	44.3	42.3	50.9
WAVE (Tang et al., 2026)	42.7	39.8	33.4	38.1	38.5
`Uni-Omni` (Contrastive baseline)	71.3	40.2	47.5	43.9	50.7
`Syn-Omni` (Ours)	71.7	41.4	**48.7**	**44.6**	**51.6**

Table 1: Comparison of omnimodal embedding models using Hit@1 (%). The numbers in parentheses indicate the number of evaluation tasks for each modality and the total. Here, `Uni-Omni` corresponds to a strong baseline trained with a standard LoRA setting. Across both 3B- and 7B-scale models, `Syn-Omni` achieves higher overall scores than previous methods, demonstrating well-balanced performance across diverse modalities, while also improving over our baseline. For each model scale, the best results are **bolded**, and the second-best are underlined.

(PSR) schedule with warmup and transition ratios of $t_{\text{warm}} = 0.1$ and $t_{\text{blend}} = 0.5$, respectively. Additional implementation details and hyperparameter configurations are provided in Sec. B.1.

Evaluation Setup. We evaluate our framework on an extensive omnimodal benchmark consisting of 81 tasks spanning image, video, audio, and audiovisual modalities. Each modality group covers multiple embedding-oriented meta-tasks, such as retrieval, classification, and QA. For image and video evaluations, we adopt the MMEB-v1 (Jiang et al., 2025) and MMEB-v2 (Meng et al., 2026) benchmarks, respectively, while further extending with additional audio and audiovisual benchmarks collected in this work. The complete evaluation dataset details are provided in Sec. A.2.

Baselines. We compare `Syn-Omni` against recent Omni-MLLM-based embedding models, including OmniEmbed-v0.1 (Ma et al., 2025a), Omni-Embed-Nemotron (Xu et al., 2025b), OmniEmbed-Multivent (Zhan et al., 2025), WAVE (Tang et al., 2026), LCO-Emb (Xiao et al., 2025), and e5-omni (Chen et al., 2026). For a fair evaluation, we use officially released checkpoints and perform all evaluations under the unified protocol.

Metrics. We use Hit@1 as the evaluation metric, measuring whether the ground-truth target is ranked at the top among candidates. We report both modality-specific average performance and the overall average score from a single run. The overall score is computed by averaging modality-specific scores, where each modality score is obtained by averaging the corresponding meta-task scores. Here, each meta-task score is computed as the average of its associated benchmark scores. Detailed benchmark-level performance breakdowns are provided in Sec. C.4. Our codes and data collections will be made publicly available.

4.1 Main Results

Balanced Performance Across Modalities. In Tab. 1, we compare our framework against recent Omni-MLLM embedding models. While some methods excel in specific modalities, such as LCO-Emb (Xiao et al., 2025) on video tasks and e5-omni (Chen et al., 2026) on image tasks, their strong performance is less consistent across modalities. In contrast, our contrastive baseline, `Uni-Omni`, trained with a modality-shared LoRA, already demonstrates strong and well-balanced performance across all modalities, achieving competitive overall scores at both the 3B and 7B scales. Building upon this strong baseline, `Syn-Omni` further improves the overall performance, reaching the best overall scores of 49.4% and 51.6% for the 3B and 7B models, respectively. The improvements are particularly evident in audio and audiovisual tasks, while maintaining competitive performance on image and video benchmarks. Overall, this demonstrates the effectiveness of our framework in balancing performance across modalities.

id	OME-LoRA S-Exp.	OME-LoRA $\mathcal{L}_{ortho}$	PSR Soft	PSR Prog.	I	V	A	AV	All
(a)	-	-	-	-	66.8	40.4	43.1	40.4	47.7
(b)	✓	-	-	-	67.6	39.6	44.9	42.6	48.7
(c)	✓	✓	-	-	67.4	39.9	45.2	41.7	48.6
(d)	✓	✓	✓	-	67.3	39.9	45.0	42.1	48.6
(e)	✓	-	✓	✓	**68.2**	40.1	45.2	42.6	49.0
(f)	✓	✓	✓	✓	68.1	**40.5**	**45.6**	**43.4**	**49.4**

Table 2: Component analysis of `Syn-Omni` framework. Starting from the `Uni-Omni` baseline in (a), each row incorporates individual components of our design. The results show that both orthogonal modality-expert LoRA and progressive synergy routing contribute to improved and more balanced performance over the baseline, with the full model achieving the best overall results.

id	Variant	Rank	PSR	I	V	A	AV	All
(a)	Baseline	16	-	66.8	40.4	43.1	40.4	47.7
(b)	Baseline	24	-	67.3	38.7	43.0	41.6	47.6
(c)	Expert-only	16	-	66.1	39.4	42.5	40.1	47.0
(d)	Shared+Expert	24	-	67.6	39.6	44.9	42.6	48.7
(e)	Expert-only	16	✓	65.3	39.5	42.1	40.5	46.9
(f)	Shared+Expert	24	✓	**68.1**	**40.5**	**45.6**	**43.4**	**49.4**

Table 3: Structural analysis of the adaptation framework. We compare our OME-LoRA architecture against single-path and expert-only adaptation variants. The results show that separating parameters into shared and expert paths improves performance more effectively than simply increasing the single-layer LoRA rank. Importantly, maintaining the shared LoRA path is essential for enabling strong performance in both routing schemes.

id	Routing Strategy (PSR) Router	Progress	loss	I	V	A	AV	All
(a)	Hard	-	-	67.4	39.9	45.2	41.7	48.6
(b)	Soft	-	BCE	67.3	39.9	45.0	42.1	48.6
(c)	Soft	-	SML	67.6	40.4	45.1	41.9	48.8
(d)	Soft	✓	BCE	67.6	39.6	**46.0**	41.9	48.8
(e)	Soft	✓	SML	**68.1**	**40.5**	45.6	**43.4**	**49.4**

Table 4: Analysis on Progressive Synergy Routing (PSR) strategy. We evaluate the impact of routing types, progressive blending, and router loss functions in the OME-LoRA layers. We highlight that combining a progressive blending of routing weights with the Synergy Margin Loss (SML) is crucial for both expert specialization and cross-modal collaboration among experts.

4.2 Ablation Study

We conduct a series of ablation experiments to analyze the individual contribution of each proposed component in our `Syn-Omni` framework. All results are reported using 3B-scale models, including both modality-specific and overall averaged scores.

Component Analysis. As shown in Tab. 2, compared to `Uni-Omni` baseline in (a), introducing a shared-expert structure (S-Exp.) with hard routing in (b) improves performance through structural decomposition. Adding the orthogonal penalty in (c) and soft routing in (d) maintain performance at a level comparable to (b), while still showing competitive results over `Uni-Omni`. When applying Progressive Synergy Routing (PSR) with Synergy Margin Loss (SML) from Eq. (6) in (e), further improvements are obtained, with our `Syn-Omni` model achieving the best overall results in (f) through the integration of the orthogonal penalty.

Structural Efficiency. We explore the impact of different architectural choices for omnimodal adaptation under a fixed low-rank parameter budget. As shown in Tab. 3, simply increasing the LoRA rank from 16 to 24 within a single shared structure in (b) fails to provide meaningful gains and even leads to a slight performance drop. In contrast, decomposing the same total rank (24) into shared and expert paths in (d) under a hard-routing scheme brings improvements. This structural advantage remains consistent when PSR is applied in (f), confirming the benefit of structured parameter decomposition.

Role of Shared LoRA. We next evaluate the necessity of the shared LoRA path within the OME-LoRA architecture. As shown in Tab. 3, omitting the shared path leads to a substantial performance drop across all tasks, under both hard routing (c vs. d) and with PSR integrated (e vs. f). Intriguingly, relying solely on modality-expert paths without a shared anchor causes performance to fall even below the `Uni-Omni` in (a). This clearly demonstrates that the shared path is essential for establishing a unified cross-modal space across heterogeneous modality inputs, thereby providing a crucial foundation for experts to develop their specialization.

Role of Progressive Synergy Routing. We further investigate the impact of our routing strategies, as summarized in Tab. 4. Compared to the hard routing baseline in (a), switching to soft routing with BCE loss in (b) provides limited to no improvement. However, across both standard soft routing (b vs. c) and progressive blending configurations (d vs. e), replacing BCE with the Synergy Margin Loss (SML) consistently improves performance. Our final framework in (e), which integrates both progressive blending and SML, achieves the best overall performance of 49.4%. This demonstrates that the two components, progressive blending with synergy margin loss in PSR complement each other to deliver the superior results.

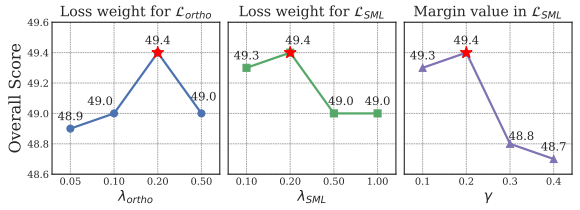

Figure 4: Sensitivity analysis of hyperparameters on the orthogonal penalty weight λ_{ortho}, the synergy margin loss weight λ_{SML}, and the margin parameter γ. Red stars indicate the optimal choice in our final model.

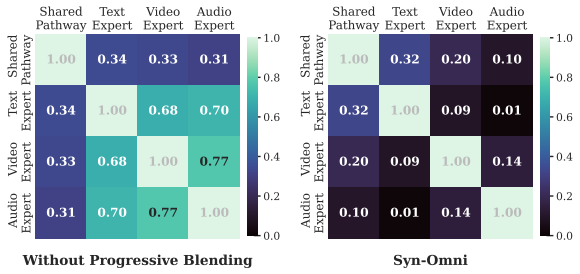

Figure 5: Pairwise cosine similarity between OME-LoRA pathways. Without progressive blending (left), high cross-path similarities reveal redundancy across different pathways. In contrast, `Syn-Omni` (right) produces significantly lower similarities, indicating effective decoupling and clear specialization across pathways.

Sensitivity Analysis. We investigate the sensitivity of three key hyperparameters in `Syn-Omni`: the orthogonal penalty weight (λ_{ortho}), the synergy margin loss weight (λ_{SML}), and the margin parameter (γ). As shown in Fig. 4, the best overall score is achieved with $\lambda_{ortho} = 0.20$, $\lambda_{SML} = 0.20$, and $\gamma = 0.2$. The performance remains relatively stable around these values, while larger margin values lead to a noticeable performance drop.

4.3 Empirical Analysis

To better understand how `Syn-Omni` structures omnimodal representations, we analyze the behaviors of its two key components, OME-LoRA and PSR, on VALOR-32k (Liu et al., 2025) evaluation samples with text, video, and audio inputs. Specifically, we analyze the decoupled parameter spaces of OME-LoRA and the expert synergy induced by PSR. Results are computed from final-layer submodule signals averaged across samples and layers.

Decoupling in OME-LoRA. As shown in Fig. 5, we measure pairwise cosine similarity between the flattened hidden states of the shared pathway and modality experts using text-video-audio inputs. Without applying progressive blending (left), relying solely on soft routing fails to induce clear separation, evidenced by high cross-path similarities both between the shared pathway and experts, and across experts. In contrast, `Syn-Omni` (right) produces substantially lower cross-path similarities, indicating clear modality specialization. These results suggest that progressive routing establishes a foundation for modality experts' specialization.

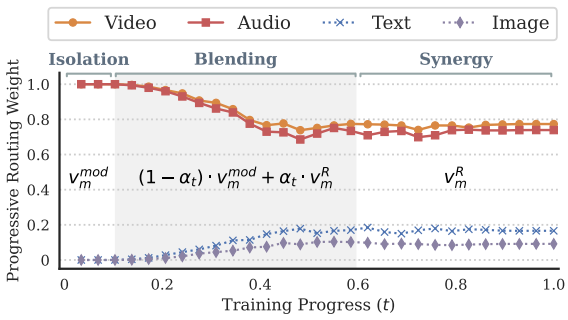

Figure 6: Routing dynamics of PSR under video-audio inputs. Present-modality experts retain dominant routing weights, while absent-modality experts gradually acquire non-zero activations through progressive blending. This indicates the emergence of synergistic cross-modal interactions along with modality specialization.

Routing Dynamics in PSR. We analyze the evolution of PSR routing weights under video-audio inputs to demonstrate the emergence of cross-modal interactions during training in Fig. 6. Under progressive blending, present-modality experts (video and audio) maintain dominant routing weights. In contrast, absent-modality experts (text and image), initially suppressed to zero under hard routing scheme, gradually acquire non-zero weights by introducing soft routing scheme. This behavior demonstrates that PSR enables cross-modal collaboration while preserving modality specialization. We further observe that SML drives this synergistic behavior, whereas BCE loss completely suppresses absent-modality experts, as shown in Fig. 8.

5 Conclusion

We propose `Syn-Omni`, a structured omnimodal adaptation framework that effectively learns a unified representation space. It leverages OME-LoRA to decouple modality-shared and specialized parameters. In addition, Progressive Synergy Routing (PSR) shifts the model from isolated specialization toward dynamic cross-modal collaboration. Across 81 diverse tasks, `Syn-Omni` consistently outperforms existing baselines, achieving robust and well-balanced performance across image, video, audio, and audiovisual domains.

Limitations

Our framework inevitably introduces additional inference overhead by dynamic routing in OME-LoRA layer. In standard LoRA layers, the adapted weights can be merged into the frozen backbone after training, eliminating inference-time overhead. In contrast, our mixture-of-LoRA design requires input-dependent routing weights at forward pass, preventing static weight merging for modality experts, although the shared pathway in OME-LoRA remains mergeable. As a result, the expert LoRA pathways remain active during inference, resulting in additional computational overhead compared to conventional LoRA-based adaptation approaches.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and 1 others. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Shuai Bai and 1 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, and 1 others. 2025. Perception encoder: The best visual embeddings are not at the output of the network. *Advances in Neural Information Processing Systems*, 38:60884–60937.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Haonan Chen, Sicheng Gao, Radu Timofte, Tetsuya Sakai, and Zhicheng Dou. 2026. e5-omni: Explicit cross-modal alignment for omni-modal embeddings. *arXiv preprint arXiv:2601.03666*.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. *URL https://arxiv. org/abs/2307.08691*, 5(10).

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. 2024. MusiLingo: Bridging music and text with pre-trained language models for music captioning and query response. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3643–3655, Mexico City, Mexico. Association for Computational Linguistics.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an audio captioning dataset. In *IEEE 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 736–740, United States. IEEE. IEEE International Conference on Acoustics, Speech and Signal Processing ; Conference date: 01-01-1900 Through 01-01-2000.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. In *International conference on machine learning*, pages 1068–1077. PMLR.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 316–321.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, and 1 others. 2017. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850.

Zhuoning Guo, Mingxin Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Xiaowen Chu. 2025. Towards universal video retrieval: Generalizing video embedding via synthesized multimodal pyramid curriculum. *arXiv preprint arXiv:2510.27571*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Mingfei Han, Linjie Yang, Xiaojun Chang, Lina Yao, and Heng Wang. 2025. Shot2story: A new benchmark for comprehensive understanding of multi-shot videos. In *International Conference on Learning Representations*, volume 2025, pages 41665–41677.

Jitai Hao, Hao Liu, Xinyan Xiao, Qiang Huang, and Jun Yu. 2026. Uni-x: Mitigating modality conflict with a two-end-separated architecture for unified multimodal models. In *The Fourteenth International Conference on Learning Representations*.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, and 1 others. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271.

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. In *International Conference on Learning Representations*.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12075.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2025. VLM2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

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.

A Sophia Koepke, Andreea-Maria Oncescu, João F Henriques, Zeynep Akata, and Samuel Albanie. 2022. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 25:2675–2685.

Hilde Kuehne, Ali Arslan, and Thomas Serre. 2014. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE.

Jie Lei, Tamara L Berg, and Mohit Bansal. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021a. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6761–6771.

Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. 2025. Valor: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):708–724.

Siqi Liu, Weixi Feng, Tsu-Jui Fu, Wenhu Chen, and William Wang. 2023. Edis: Entity-driven image search over multimodal web content. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4877–4894.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021b. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2125–2134.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521.

Xueguang Ma, Luyu Gao, Shengyao Zhuang, Jiaqi Samantha Zhan, Jamie Callan, and Jimmy Lin. 2025a. Tevatron 2.0: Unified document retrieval toolkit across scale, language, and modality. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4061–4065.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024a. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505.

Yufei Ma, Zihan Liang, Huangyu Dai, Ben Chen, Dehong Gao, Zhuoran Ran, Wang Zihan, Linbo Jin, Wen Jiang, Guannan Zhang, Xiaoyan Cai, and Libin Yang. 2024b. MoDULA: Mixture of domain-specific and universal LoRA for multi-task learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2758–2770, Miami, Florida, USA. Association for Computational Linguistics.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, and 15 others. 2025b. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, volume 36, pages 46212–46244. Curran Associates, Inc.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

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.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2024. Mustango: Toward controllable text-to-music generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8293–8316.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Raghuveer Thirukovalluru, Xuan Zhang, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. 2026. VLM2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *Transactions on Machine Learning Research*.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576.

Karol J Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018.

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

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pages 84929–84964.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. 2025. AVHBench: A cross-modal hallucination benchmark for audio-visual large language models. In *The Thirteenth International Conference on Learning Representations*.

Changli Tang, Qinfan Xiao, Ke Mei, Tianyi Wang, Fengyun Rao, and Chao Zhang. 2026. WAVE: Learning unified & versatile audio-visual embeddings with multimodal LLM. In *The Fourteenth International Conference on Learning Representations*.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263.

George Tzanetakis and Perry Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, and 1 others. 2024. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pages 396–416. Springer.

Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. 2022. N24news: A new dataset for multimodal news classification. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6768–6775.

Shicai Wei, Chunbo Luo, and Yang Luo. 2025. Boosting multimodal learning via disentangled gradient learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22879–22888.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317.

Qiyu Wu, Shuyang Cui, Satoshi Hayakawa, Wei-Yao Wang, Hiromi Wakaki, and Yuki Mitsufuji. 2025. Mca: Modality composition awareness for robust composed multimodal retrieval. *arXiv preprint arXiv:2510.15543*.

Yue WU, Zhaobo Qi, Yiling Wu, Junshu Sun, Yaowei Wang, and Shuhui Wang. 2025. Learning fine-grained representations through textual token disentanglement in composed video retrieval. In *The Thirteenth International Conference on Learning Representations*.

Chenghao Xiao, Hou Pong Chan, Hao Zhang, Weiwen Xu, Mahani Aljunied, and Yu Rong. 2025. Scaling language-centric omnimodal representation learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786.

Jilan Xu, Carl Thomé, Danijela Horak, Weidi Xie, and Andrew Zisserman. 2026. Scaling audio-text retrieval with multimodal large language models. *arXiv preprint arXiv:2602.18010*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, and 1 others. 2025a. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

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. 2025b. Omni-embed-nemotron: A unified multimodal retrieval model for text, image, audio, and video. *arXiv preprint arXiv:2510.03458*.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3480–3491.

Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. 2024. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *European Conference on Computer Vision*, pages 146–164. Springer.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: a dataset for understanding complex web videos via question answering. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Huaying Yuan, Jian Ni, Zheng Liu, Yueze Wang, Junjie Zhou, Zhengyang Liang, Bo Zhao, Zhao Cao, Ji-Rong Wen, and Zhicheng Dou. 2025. Momentseeker: A task-oriented benchmark for long-video moment retrieval. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc.

Jiaqi Samantha Zhan, Crystina Zhang, Shengyao Zhuang, Xueguang Ma, and Jimmy Lin. 2025. Magmar shared task system description: Video retrieval with omniembed. *arXiv preprint arXiv:2506.09409*.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander G Hauptmann, Yonatan Bisk, and 1 others. 2025. Direct preference optimization of video large multimodal models from language model reward. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 694–717.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. 2025. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

Contents

A Additional Details **14**
A.1 Training Datasets 14
A.2 Evaluation Datasets 14

B Detailed Experimental Settings **15**
B.1 Training Details & Hyperparameters 15
B.2 Evaluation Pipeline 15

C Detailed Results **16**
C.1 Routing Behavior across Layers . 16
C.2 Modality-Specific Models 16
C.3 Runtime Analysis 16
C.4 Per-dataset Evaluation 17

D AI Tool Usage **17**

E Artifact **17**

F Ethical Considerations **21**

A Additional Details

A.1 Training Datasets

To facilitate omnimodal learning, we construct a comprehensive training mixture categorized into four modality groups: image, video, audio, and audiovisual. Beyond single modalities, this mixture features composed modality queries and/or targets, allowing the model to learn versatile alignment spaces. The distribution and coverage of sample counts for each query-to-target modality composition are visualized in Fig. 7, while detailed dataset-level specifications for each modality group are provided in Tabs. 8 to 11, respectively. For each dataset, if the original dataset size exceeded our specified target size, we randomly sub-sampled it to match that specified size.

As shown in Fig. 1, while the modality-specific models were trained independently only on the datasets corresponding to their respective modality groups, our `Uni-Omni` and `Syn-Omni` models were jointly trained on the entire mixture encompassing all modalities simultaneously.

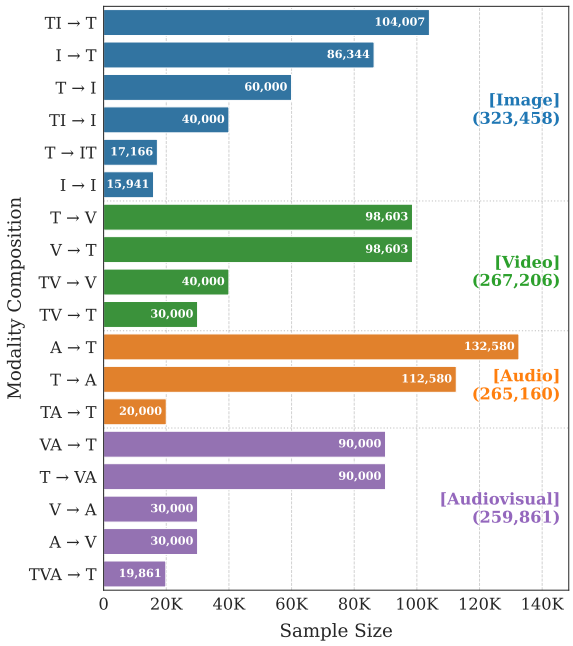

Figure 7: Distribution of sample counts for each query-to-target modality composition in our training mixture. For each composition, 'T', 'I', 'V', and 'A' correspond to text, image, video, and audio modality, respectively. We also specify the total sample size per modality group.

A.2 Evaluation Datasets

To construct our evaluation suite across four distinct modalities: image, video, audio, and audiovisual, we leverage and extend existing evaluation frameworks. Specifically, for the image and video tasks, we utilize the evaluation suites from MMEB-v1 (Jiang et al., 2025) and MMEB-v2 (Meng et al., 2026), respectively. While MMEB-v2 natively includes text-to-video retrieval, it lacks the corresponding video-to-text direction. Therefore, we integrated video-to-text retrieval to ensure a bidirectional evaluation. Furthermore, we expand the benchmark's scope by incorporating additional audio and audiovisual tasks.

The detailed statistics and configurations for each modality are provided from Tab. 12 to Tab. 15, each specifying the task type, query-to-target composition, and data sizes. In these tables, the corpus size is defined differently based on the task type: it denotes the number of classes for classification (CLS) tasks, the number of choices for question answering (QA) tasks, and the actual database size for retrieval (RET) tasks. Our data collection includes a wide variety of common-domain settings for each modality, such as integrating both generic audio and music domain. English is the primary language across all multi-modal tasks.

B Detailed Experimental Settings

B.1 Training Details & Hyperparameters

Model Backbone. We select the pretrained Qwen-2.5-Omni (Xu et al., 2025a) as the backbone. During training, the visual and audio encoders are frozen, and LoRA parameters are inserted into the LLM layers. The LoRA target modules are `q_proj`, `k_proj`, `v_proj`, `o_proj`, `down_proj`, `up_proj`, and `gate_proj`, which are applied to each of the 36 LLM layers of the Qwen-2.5-Omni backbone.

Input Processing. For the input processing, we adopt the native configuration of Qwen-2.5-Omni (Xu et al., 2025a) unless specified otherwise. Specifically, images and videos are processed using the native resolution scaling strategy of Qwen-2.5-VL (Bai et al., 2025). The minimum and maximum image resolutions are set to 50,176 pixels ($64 \times 28 \times 28$) and 784,000 pixels ($1000 \times 28 \times 28$), respectively. For video inputs, videos are uniformly sampled into 8 frames, and the per-frame resolution is bounded between a minimum of 100,352 pixels ($128 \times 28 \times 28$) and a maximum of 156,800 pixels ($200 \times 28 \times 28$). Audio signals are sampled at 16,000 Hz. The textual inputs are truncated to a maximum of 512 text tokens (without chat templates), and consistent with prior approaches (Meng et al., 2026; Chen et al., 2026), left-padding is applied during batching.

LoRA Configurations. For the modality-specific models and the standard LoRA baseline `Uni-Omni` mentioned in Fig. 1, the LoRA rank (r) is 16. For `Syn-Omni` with OME-LoRA, the rank of the shared adaptation path is 8, and each modality expert rank is 4, making the total effective rank 24. Across all configurations, the LoRA scaling factor α is set to two times the rank. LoRA dropout is applied with probability 0.1. The detailed parameter counts for both the `Uni-Omni` baseline and `Syn-Omni` variants under these LoRA configurations are summarized in Tab. 5. Notably, the modality router introduces minimal parameter overhead, less than 8% of the total trainable parameters.

Optimization. All models are trained for one epoch using the Hugging Face Trainer with FlashAttention-2 (Dao, 2023) in BFloat16 mixed-precision. We use the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We employ a cosine learning rate schedule with a peak learning rate of 10^{-4}, a warmup ratio of 0.04, a weight decay of 0.01, and a maximum gradient norm of 1.0. The global batch size is set to 768. We employ GradCache (Gao et al., 2021) to increase the global batch size, where the chunk size is set to 1 for the 3B models and 4 for the 7B models. We do not employ hard negative examples. The 3B models are trained using 8× NVIDIA A100 (40GB) GPUs, which takes approximately 5 days for `Syn-Omni`, and the 7B models are trained using 8× NVIDIA H200 (141GB) GPUs, taking approximately 40 hours for `Syn-Omni`.

	`Uni-Omni`		`Syn-Omni`	
	3B	7B	3B	7B
Backbone	Qwen-2.5-Omni			
LoRA rank (Shared)	16		8	
LoRA rank (Expert)	0		4×4	
LoRA rank (Total)	16		24	
LoRA params.	29.9 M	40.4 M	44.9 M	60.6 M
Router params.	0.0 M	0.0 M	3.4 M	4.5 M
Total trainable params.	29.9 M	40.4 M	48.3 M	65.1 M

Table 5: Summary of LoRA configuration and trainable parameters for `Uni-Omni` and `Syn-Omni` model variants.

Training Objective. The baseline models and modality-specific models are optimized using the contrastive loss. `Syn-Omni` is optimized using the contrastive loss combined with the orthogonal penalty and the synergy margin loss as proposed in the main paper. The contrastive loss uses a temperature parameter τ of 0.02. For the orthogonal penalty, the loss weight is set to 0.2. The Progressive Synergy Routing (PSR) schedule uses warmup and transition ratios of $t_{\text{warm}} = 0.1$ and $t_{\text{blend}} = 0.5$. For the synergy margin loss, the margin γ is 0.2 and the loss weight is set to 0.2.

B.2 Evaluation Pipeline

For a fair evaluation, previous methods are evaluated under the unified setting using the official checkpoints. We extend the VLM2Vec-V2 codebase used for evaluation (Meng et al., 2026), which enables the integrated evaluation of diverse modalities and task types while facilitating seamless modality expansion. While this evaluation framework originally supports image and video inputs, we extend it to encompass audio and audiovisual modalities also, thereby enabling an any-to-any query-to-target retrieval setting.

The input processing for text, image, video, and audio modalities remains identical to those used during the training phase. For inference, standard LoRA adapter parameters are merged into the base pretrained weights, and such LoRA layers are not

utilized any further. In contrast, for Syn-Omni, only the shared adaptation path is merged into the backbone weights and removed, whereas the expert paths are explicitly maintained to support dynamic routing during inference, where the soft routing weights generated by the router are utilized.

C Detailed Results

C.1 Routing Behavior across Layers

Extending the analysis in Fig. 6, we investigate how different router loss formulations affect layer-level routing dynamics. While the main paper tracks routing weights across training steps, this section visualizes the layer-wise soft routing activations of the converged model using audiovisual, video-only, and audio-only inputs from VALOR-32k (Liu et al., 2025). Under this configuration, Fig. 8 compares the binary cross-entropy (BCE) loss against our synergy margin loss across the entire layer hierarchy.

As shown in the figure, BCE loss forces the router into strict modality classification based on the input type. Except for the first layer, this results in routing activations approaching 1 for present modalities and 0 for absent ones, collapsing into a deterministic hard routing behavior despite using a soft routing formulation. In contrast, our synergy margin loss maintains dominant activations for the input modalities while preserving meaningful activations even for absent modalities across all layers. This collaborative activation behavior across the layer hierarchy successfully facilitates cross-modal synergy, aligning with our architectural objectives.

C.2 Modality-Specific Models

We present the extended evaluation of the modality-specific models introduced in Fig. 1, reporting their performance across all modalities in Tab. 6. Models trained exclusively on a single modality demonstrate strong performance within their respective modality. For instance, the audio-specific model performs well on audio tasks. A notable exception is observed in the audiovisual-specific model. This indicates that exposure to individual image, video, and audio data within a unified training scheme is helpful for audiovisual tasks.

Intriguingly, the image-specific model exhibits solid performance on video tasks, while the video-specific model retains reasonable results on both image and audiovisual tasks. These trends indicate that a unified training setting, where a model is exposed to diverse modalities, can induce cross-modal synergy. This serves as the primary motivation behind the design of our Orthogonal Modality-Expert LoRA (OME-LoRA) paired with the Progressive Synergy Routing (PSR) mechanism.

	I	V	A	AV	All
I-specific	68.0	43.0	35.4	33.3	44.9
V-specific	48.1	41.1	34.9	35.9	40.0
A-specific	40.3	36.5	43.5	30.9	37.8
AV-specific	38.0	32.4	34.6	37.1	35.5
Uni-Omni (3B)	66.8	40.4	43.1	40.4	47.7
Syn-Omni (3B)	68.1	40.5	45.6	43.4	49.4

Table 6: Comprehensive evaluation of modality-specific models on all modalities. The results demonstrate intrinsic cross-modal transferability and highlight the necessity of unified multimodal training for cross-modal synergy. Here, I, V, A, and AV denote image, video, audio, and audiovisual modalities, respectively.

Model	Latency (ms)	Peak GPU Mem. (MB)
Uni-Omni (3B)	300.3 ± 25	9612
Syn-Omni (3B)	354.6 ± 20	9676
Uni-Omni (7B)	377.1 ± 26	17672
Syn-Omni (7B)	443.5 ± 30	17765

Table 7: Runtime analysis on Uni-Omni and our Syn-Omni model variants. We report mean latency (ms) and peak GPU memory usage (MB), testing on VALOR-32k audiovisual input samples.

While the Uni-Omni baseline exhibits performance comparable to or slightly lower than that of the respective modality-specific models in their native modalities, our Syn-Omni framework consistently delivers improved performance, demonstrating the effectiveness of cross-modal synergy.

C.3 Runtime Analysis

To evaluate inference-time efficiency, we benchmark Uni-Omni and Syn-Omni on the VALOR-32k audiovisual input using a single NVIDIA RTX A6000 GPU (batch size 1). As shown in Tab. 7, Syn-Omni introduces a slight increase in both latency and memory across 3B and 7B scales. Specifically, for the 3B variant, the mean latency increases from 300.3 ms to 354.6 ms, while the peak GPU memory changes from 9612 MB to 9676 MB.

This overhead is due to architectural differences at inference time. In Uni-Omni, standard LoRA adapters are fully merged into the backbone weights, maintaining the original backbone's infer-

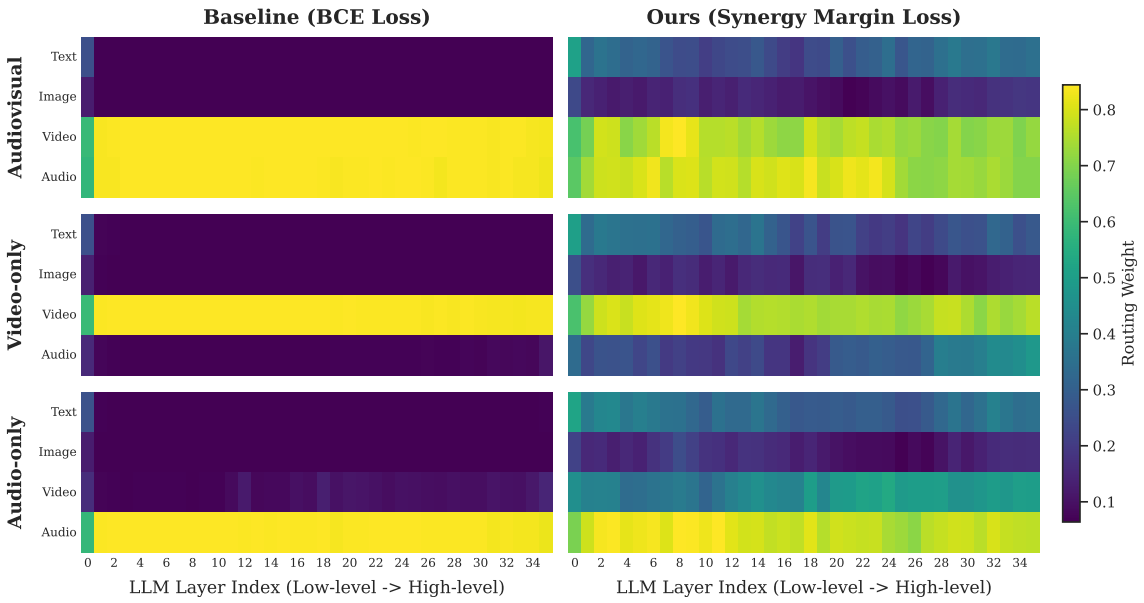

Figure 8: Visualization of layer-wise soft routing weights under different input modalities. The baseline (BCE loss) exhibits a collapsed hard routing behavior with nearly binary activations, whereas ours (synergy margin loss) dynamically allocates routing weights even to absent-modality experts to encourage cross-modal synergy.

ence path. In contrast, while `Syn-Omni` merges its shared adaptation path, it explicitly retains the four modality-specific expert paths and router parameters to support input-dependent routing.

C.4 Per-dataset Evaluation

We present the comprehensive benchmark-level results across all evaluated modalities and tasks. Individual dataset performances for image, video, audio, and audiovisual tasks are detailed in Tab. 16, Tab. 17, Tab. 18, and Tab. 19, respectively.

D AI Tool Usage

We utilized ChatGPT to improve the writing fluency and grammar of the manuscript. The development of the core ideas, technical methodologies, experimental designs, and analyses remains the original work of the authors.

E Artifact

We meticulously document the licenses, access URLs, and intended uses of the artifacts utilized in this study. All datasets, models, and software artifacts were used in accordance with their original licenses and intended research usage conditions when specified. For redistributed repositories or mirrors, we additionally verified the original artifact licenses when available. All derivatives or model checkpoints in our work are targeted for non-commercial research purposes.

- **MMEB v1 Training Set** (Jiang et al., 2025): Sourced from the official HuggingFace repository[1] under the `Apache-2.0` license.
- **LLaVAHound** (Zhang et al., 2025): Sourced from the official HuggingFace repository[2] under the `Apache-2.0` license.
- **PE-Video** (Bolya et al., 2025): Sourced from the official HuggingFace repository[3] under the `CC-BY-NC-4.0` license for non-commercial academic research.
- **MSVD** (Chen and Dolan, 2011) & **MSR-VTT** (Xu et al., 2016): Sourced from a public redistributed HuggingFace repository[4][5] while original licenses unspecified.
- **ActivityNet Captions** (Caba Heilbron et al., 2015): Sourced from a redistributed HuggingFace repository[6] under the `Apache License 2.0`.

[1] https://huggingface.co/datasets/TIGER-Lab/MMEB-train
[2] https://huggingface.co/datasets/ShareGPTVideo/train_video_and_instruction
[3] https://huggingface.co/datasets/facebook/PE-Video
[4] https://huggingface.co/datasets/friedrichor/MSVD
[5] https://huggingface.co/datasets/friedrichor/MSR-VTT
[6] https://huggingface.co/datasets/friedrichor/ActivityNet_Captions

Dataset	Composition	Size
ImageNet 1K (Deng et al., 2009)	I → T	15,000
N24News (Wang et al., 2022)	TI → T	15,000
HatefulMemes (Kiela et al., 2020)	I → T	8,500
VOC2007 (Everingham et al., 2015)	I → T	7,844
SUN397 (Xiao et al., 2010)	I → T	15,000
OK-VQA (Marino et al., 2019)	TI → T	9,007
A-OKVQA (Schwenk et al., 2022)	TI → T	15,000
DocVQA (Mathew et al., 2021)	TI → T	15,000
InfographicsVQA (Mathew et al., 2022)	TI → T	15,000
ChartQA (Masry et al., 2022)	TI → T	15,000
Visual7W (Zhu et al., 2016)	TI → T	20,000
VisDial (Das et al., 2017)	T → I	20,000
CIRR (Liu et al., 2021b)	TI → I	20,000
VisualNews (Liu et al., 2021a)	T → I	20,000
	I → T	20,000
MSCOCO-Retrieval (Lin et al., 2014)	T → I	20,000
	I → T	20,000
MSCOCO-Grounding (Lin et al., 2014)	TI → I	20,000
NIGHTS (Fu et al., 2023)	I → I	15,941
WebQA (Chang et al., 2022)	T → IT	17,166
Sum		**323,458**

Table 8: An overview of image-text datasets in the omnimodal training dataset. Each dataset is sourced from the MMEB-v1 (Jiang et al., 2025), where either the full original set or a randomly sampled subset was utilized.

- **DiDeMo** (Anne Hendricks et al., 2017): Sourced from a redistributed HuggingFace repository[7] under the `BSD 2-Clause` license.
- **FineCVR** (WU et al., 2025): Sourced from the official GitHub repository[8], while license information is unknown.
- **AudioCaps** (Kim et al., 2019): Sourced from a redistributed HuggingFace mirror[9] under the `MIT` license.
- **WavCaps** (Mei et al., 2024): Sourced from the official HuggingFace repository[10] under the `CC-BY-4.0` license.
- **AudioSet Strong** (Gemmeke et al., 2017): Sourced from a redistributed HuggingFace mirror[11] under the `CC BY 4.0` specification.
- **MusicCaps** (Agostinelli et al., 2023): Sourced from a redistributed HuggingFace mirror[12] under the `CC-BY-SA-4.0` framework.
- **MusicBench** (Melechovsky et al., 2024): Sourced from the official HuggingFace repository[13] under the `CC-BY-SA-3.0` license.
- **MusicInstruct** (Deng et al., 2024): Sourced from the official HuggingFace repository[14] under the `CC-BY-NC-4.0` license.
- **AudioSet** (Gemmeke et al., 2017): Sourced from the official page[15] under the `CC BY 4.0` license.
- **VAST** (Chen et al., 2023): Sourced from the official GitHub repository[16] under the `MIT` license.

[7] https://huggingface.co/datasets/friedrichor/DiDeMo
[8] https://github.com/May2333/FDCA
[9] https://huggingface.co/datasets/CLAPv2/audiocaps
[10] https://huggingface.co/datasets/cvssp/WavCaps
[11] https://huggingface.co/datasets/CLAPv2/audioset_strong
[12] https://huggingface.co/datasets/CLAPv2/MusicCaps
[13] https://huggingface.co/datasets/amaai-lab/MusicBench
[14] https://huggingface.co/datasets/m-a-p/Music-Instruct
[15] https://research.google.com/audioset/
[16] https://github.com/CASIA-IVA-Lab/VAST

Dataset	Composition	Size
LLaVAHound Retrieval (Zhang et al., 2025)	T → V V → T	30,000 30,000
LLaVAHound QA (Zhang et al., 2025)	TV → T	30,000
PE-Video (Bolya et al., 2025)	V → T T → V	40,000 40,000
MSVD (Chen and Dolan, 2011)	V → T T → V	1,200 1,200
MSR-VTT (Xu et al., 2016)	V → T T → V	9,000 9,000
ActivityNet (Caba Heilbron et al., 2015)	V → T T → V	10,008 10,008
DiDeMo (Anne Hendricks et al., 2017)	V → T T → V	8,395 8,395
FineCVR (WU et al., 2025)	TV → V	40,000
Sum		**267,206**

Table 9: An overview of video-centric datasets included in the omnimodal training dataset.

- **InternVideo2** (Wang et al., 2024): Sourced from the official HuggingFace repository[17] under the `CC-BY-NC-SA-4.0` license.
- **Shot2Story** (Han et al., 2025): Sourced from the official HuggingFace repository[18] under the `CC-BY-NC-SA-4.0` license, restricted to non-commercial academic usage.
- **AVInstruct** (Ye et al., 2024): Sourced from the official GitHub repository[19] under the `Apache-2.0` license, utilizing video sources derived from AVQA (Yang et al., 2022).
- **MMEB v1 Evaluation Set** (Jiang et al., 2025): Sourced from the official HuggingFace repository[20] under the `Apache-2.0` license.
- **MMEB v2 Evaluation Set** (Meng et al., 2026): Sourced from the official HuggingFace repository[21] under the `Apache-2.0` protocol.
- **ESC-50** (Piczak, 2015): Sourced from a redistributed HuggingFace mirror[22] under the `CC-BY-NC`.
- **GTZAN** (Tzanetakis and Cook, 2002): Sourced from a redistributed HuggingFace mirror[23].
- **NSynth** (Engel et al., 2017): Sourced from a redistributed HuggingFace mirror[24] under the `CC BY 4.0` license.
- **AudioCaps (for Eval)** (Kim et al., 2019): Sourced from a redistributed HuggingFace mirror[25] under the `MIT` license.
- **Clotho** (Drossos et al., 2020): Sourced from the official archive [26] under its custom academic non-commercial research license.
- **SoundDescs** (Koepke et al., 2022) & **AVE** (Tian et al., 2018): Sourced from a redistributed HuggingFace repository[27]. **SoundDescs** follows the `Apache License 2.0`, while the license information for **AVE** was not explicitly provided in the original repository.
- **MMAU** (Sakshi et al., 2025): Sourced from a

[17] https://huggingface.co/datasets/OpenGVLab/InternVideo2_Vid_Text
[18] https://huggingface.co/ByteDance/shot2story
[19] https://github.com/rikeilong/Bay-CAT
[20] https://huggingface.co/datasets/TIGER-Lab/MMEB-eval
[21] https://huggingface.co/datasets/TIGER-Lab/MMEB-V2
[22] https://huggingface.co/datasets/CLAPv2/esc50_no_overlap
[23] https://huggingface.co/datasets/m-a-p/GTZAN
[24] https://huggingface.co/datasets/TwinkStart/Nsynth
[25] https://huggingface.co/datasets/OpenSound/AudioCaps
[26] https://zenodo.org/records/4783391
[27] https://huggingface.co/datasets/VLM2Vec/MMEB-V3

Dataset	Composition	Size
AudioCaps (Kim et al., 2019)	A → T T → A	45,000 45,000
WavCaps (Mei et al., 2024)	A → T T → A	45,000 45,000
AudioSet-SL (Gemmeke et al., 2017)	A → T	20,000
MusicCaps (Agostinelli et al., 2023)	A → T T → A	2,580 2,580
MusicBench (Melechovsky et al., 2024)	A → T T → A	20,000 20,000
MusicInstruct (Deng et al., 2024)	TA → T	20,000
Sum		**265,160**

Table 10: An overview of audio-centric datasets included in the omnimodal training dataset.

Dataset	Composition	Size
AudioSet (Gemmeke et al., 2017)	V → A A → V	30,000 30,000
VAST (Chen et al., 2023)	VA → T T → VA	30,000 30,000
InternVideo2 (Wang et al., 2024)	VA → T T → VA	30,000 30,000
Shot2Story (Han et al., 2025)	VA → T T → VA	30,000 30,000
AVInstruct (Ye et al., 2024)	TVA → T	19,861
Sum		**259,861**

Table 11: An overview of audiovisual-centric datasets included in the omnimodal training dataset.

redistributed HuggingFace repository[28] under the `CC-BY-SA-4.0` standards.

- **MMAR** (Ma et al., 2025b): Sourced from the official HuggingFace repository[29] under the `CC-BY-NC-4.0` license.
- **VALOR 32k** (Liu et al., 2025): Sourced from the official GitHub repository[30] under the `MIT` framework.
- **AVHBench** (Sung-Bin et al., 2025): Sourced from the official GitHub repository[31].
- **DailyOmni** (Zhou et al., 2025): Sourced from a redistributed HuggingFace repository[32] under the `GPL-3.0` license.
- **Qwen 2.5 Omni** (Xu et al., 2025a): Sourced from the official HuggingFace repository[33] under the `Apache-2.0` license.
- **Omni-Embed-Nemotron** (Xu et al., 2025b): Sourced from the official HuggingFace reposi-

[28] https://huggingface.co/datasets/gamma-lab-umd/MMAU-test-mini
[29] https://huggingface.co/datasets/BoJack/MMAR
[30] https://github.com/CASIA-IVA-Lab/VALOR
[31] https://github.com/kaist-ami/AVHBench
[32] https://huggingface.co/datasets/mteb/Daily-Omni
[33] https://huggingface.co/Qwen/Qwen2.5-Omni-7B

tory[34] under NVIDIA's custom non-commercial scientific license (`customized-nscl-v1`) for non-commercial research and scientific evaluation.

- **LCO-Emb** (Xiao et al., 2025): Sourced from the official HuggingFace repository[35] under the `Apache-2.0` policy.

- **e5-omni** (Chen et al., 2026): Sourced from the official HuggingFace repository[36] under the `MIT` license.

- **OmniEmbed-v0.1** & **OmniEmbed-Multivent** (Ma et al., 2025a; Zhan et al., 2025): Sourced from the official HuggingFace repository[37] [38] under the `MIT` license.

- **WAVE** (Tang et al., 2026): Sourced from the official GitHub repository[39] under the `Apache License 2.0`.

- **GradCache** (Gao et al., 2021): Sourced from the official GitHub repository[40] under the `Apache-2.0` license.

- **Flash-Attention-2** (Dao, 2023): Sourced from the official GitHub repository[41] under the `BSD-3-Clause` license.

- **qwen-omni-utils**: Sourced from the official GitHub repository[42] under the `Apache-2.0` license.

- **Tevatron 2.0** (Ma et al., 2025a): Sourced from the official GitHub repository[43] under the `Apache-2.0` protocol.

All datasets were obtained from public benchmarks and repositories, and we relied on the original providers' release procedures for potentially identifiable or offensive content.

[34] https://huggingface.co/nvidia/omni-embed-nemotron-3b
[35] https://huggingface.co/LCO-Embedding
[36] https://huggingface.co/collections/Haon-Chen/e5-omni
[37] https://huggingface.co/Tevatron/OmniEmbed-v0.1
[38] https://huggingface.co/Tevatron/OmniEmbed-v0.1-multivent
[39] https://github.com/TCL606/WAVE
[40] https://github.com/luyug/GradCache
[41] https://github.com/dao-ailab/flash-attention
[42] https://github.com/qwenlm/qwen2.5-omni
[43] https://github.com/texttron/tevatron

F Ethical Considerations

Our framework may inherit systemic biases from large-scale training data, leading to skewed cross-modal retrieval behaviors. Additionally, such unified omnimodal embeddings could be misused for unauthorized surveillance or identity linkage, necessitating strict deployment oversight.

Task	Dataset	Composition	Query Size	Corpus Size
I-CLS	ImageNet-1K (Deng et al., 2009)	I2T	1000	1000
	N24News (Wang et al., 2022)	TI2T	1000	24
	HatefulMemes (Kiela et al., 2020)	I2T	1000	2
	VOC2007 (Everingham et al., 2015)	I2T	1000	20
	SUN397 (Xiao et al., 2010)	I2T	1000	397
	Place365 (Zhou et al., 2017)	I2T	1000	365
	ImageNet-A (Hendrycks et al., 2021b)	I2T	1000	1000
	ImageNet-R (Hendrycks et al., 2021a)	I2T	1000	200
	ObjectNet (Barbu et al., 2019)	I2T	1000	113
	Country211 (Radford et al., 2021)	I2T	1000	211
I-QA	OK-VQA (Marino et al., 2019)	TI2T	1000	1000
	A-OKVQA (Schwenk et al., 2022)	TI2T	1000	894
	DocVQA (Mathew et al., 2021)	TI2T	1000	1000
	InfographicsVQA (Mathew et al., 2022)	TI2T	1000	1000
	ChartQA (Masry et al., 2022)	TI2T	1000	1000
	Visual7W (Zhu et al., 2016)	TI2T	1000	1000
	ScienceQA (Lu et al., 2022)	TI2T	1000	736
	VizWiz (Gurari et al., 2018)	TI2T	1000	1000
	GQA (Hudson and Manning, 2019)	TI2T	1000	1000
	TextVQA (Singh et al., 2019)	TI2T	1000	1000
IT RET	MSCOCO_i2t (Lin et al., 2014)	I2T	1000	1000
	VisualNews_i2t (Liu et al., 2021a)	I2T	1000	1000
	VisDial (Das et al., 2017)	T2I	1000	1000
	MSCOCO_t2i (Lin et al., 2014)	T2I	1000	1000
	VisualNews_t2i (Liu et al., 2021a)	T2I	1000	1000
	WebQA (Chang et al., 2022)	T2TI	1000	1000
	EDIS (Liu et al., 2023)	T2TI	1000	1000
	Wiki-SS-NQ (Ma et al., 2024a)	T2I	1000	1000
	CIRR (Liu et al., 2021b)	TI2I	1000	1000
	NIGHTS (Fu et al., 2023)	I2I	1000	1000
	OVEN (Hu et al., 2023)	TI2TI	1000	1000
	FashionIQ (Wu et al., 2021)	TI2I	1000	1000
VG	MSCOCO (Lin et al., 2014)	TI2I	1000	1000
	RefCOCO (Kazemzadeh et al., 2014)	TI2I	1000	1000
	RefCOCO-Matching (Kazemzadeh et al., 2014)	TI2TI	1000	1000
	Visual7W-Pointing (Zhu et al., 2016)	TI2I	1000	1000

Table 12: Evaluation datasets for image modality, including task type and query-to-target composition.

Task	Dataset	Composition	Query Size	Corpus Size
V-CLS	SmthSmthV2 (Goyal et al., 2017)	V2T	1000	174
	HMDB51 (Kuehne et al., 2011)	V2T	1000	51
	UCF101 (Soomro et al., 2012)	V2T	1000	101
	K700 (Carreira et al., 2019)	V2T	1000	700
	Breakfast (Kuehne et al., 2014)	V2T	433	10
T2V RET	MSR-VTT (Xu et al., 2016)	T2V	1000	1000
	MSVD (Chen and Dolan, 2011)	T2V	670	670
	DiDeMo (Anne Hendricks et al., 2017)	T2V	1004	1004
	YouCook2 (Zhou et al., 2018)	T2V	3179	3179
	VATEX (Wang et al., 2019)	T2V	4478	4478
V2T RET	MSR-VTT (Xu et al., 2016)	V2T	1000	995
	MSVD (Chen and Dolan, 2011)	V2T	670	660
	DiDeMo (Anne Hendricks et al., 2017)	V2T	1004	1004
	YouCook2 (Zhou et al., 2018)	V2T	3179	3105
	VATEX (Wang et al., 2019)	V2T	4478	4478
M-RET	QVHighlight (Lei et al., 2021)	TV2V	1083	10
	Charades-STA (Gao et al., 2017)	TV2V	727	10
	MomentSeeker (Yuan et al., 2025)	TV2V	1602	10
V-VQA	Video-MME (Fu et al., 2025)	TV2T	2700	4
	NExTQA (Xiao et al., 2021)	TV2T	8564	5
	EgoSchema (Mangalam et al., 2023)	TV2T	500	5
	MVBench (Li et al., 2024)	TV2T	4000	5
	ActivityNetQA (Yu et al., 2019)	TV2T	1000	2

Table 13: Evaluation datasets for video modality, including task type and query-to-target composition.

Task	Dataset	Composition	Query Size	Corpus Size
A-CLS	ESC50 (Piczak, 2015)	A2T	1237	50
	GTZAN (Tzanetakis and Cook, 2002)	A2T	290	10
	NSynth (Engel et al., 2017)	A2T	4096	10
T2A RET	AudioCaps (Kim et al., 2019)	T2A	4411	883
	Clotho (Drossos et al., 2020)	T2A	5225	1045
	MusicCaps (Agostinelli et al., 2023)	T2A	2772	2772
	SoundDescs (Koepke et al., 2022)	T2A	1000	1000
A2T RET	AudioCaps (Kim et al., 2019)	A2T	883	4201
	Clotho (Drossos et al., 2020)	A2T	1045	5225
	MusicCaps (Agostinelli et al., 2023)	A2T	2772	2772
A-QA	MMAU (Sakshi et al., 2025)	TA2T	1000	4
	MMAR (Ma et al., 2025b)	TA2T	995	4

Table 14: Evaluation datasets for audio modality, including task type and query-to-target composition.

Task	Dataset	Composition	Query Size	Corpus Size
A2V RET	AVE (Tian et al., 2018)	A2V	402	402
	VALOR32k (Liu et al., 2025)	A2V	3239	3239
V2A RET	AVE (Tian et al., 2018)	V2A	402	402
	VALOR32k (Liu et al., 2025)	V2A	3239	3239
T2VA RET	AVHBench (Sung-Bin et al., 2025)	T2VA	1105	1105
	VALOR32k (Liu et al., 2025)	T2VA	3239	3239
VA2T RET	AVHBench (Sung-Bin et al., 2025)	VA2T	1105	1105
	VALOR32k (Liu et al., 2025)	VA2T	3239	3239
AV-QA	AVHBenchQA (Sung-Bin et al., 2025)	TVA2T	5293	2
	DailyOmni (Zhou et al., 2025)	TVA2T	1196	4

Table 15: Evaluation datasets for audiovisual modality, including task type and query-to-target composition.

	Nemotron (3B)	LCO-Emb (3B)	e5-omni (3B)	Uni-Omni (3B)	Syn-Omni (7B)	OmniEmbed (7B)	Multivent (7B)	LCO Emb (7B)	e5-omni (7B)	WAVE (7B)	Uni-Omni (7B)	Syn-Omni (7B)
Average												
Overall (36)	44.1	58.1	64.8	66.8	68.1	44.2	51.8	61.6	72.5	42.7	71.3	71.7
I-CLS (10)	47.9	57.3	60.7	64.8	65.4	44.5	56.5	59.5	67.6	49.0	66.1	67.5
I-QA (10)	20.1	58.2	61.3	63.1	62.7	22.5	29.6	62.7	69.9	25.8	67.8	66.8
I-RET (12)	58.7	55.4	68.3	66.3	66.3	49.5	61.1	58.9	71.9	44.1	68.3	69.1
VG (4)	49.5	61.5	69.1	72.8	77.8	60.5	60.0	65.1	80.9	51.9	83.2	83.6
Image Classification (I-CLS)												
ImageNet-1K	60.6	65.5	72.9	74.2	75.0	56.4	68.6	72.0	79.6	47.8	76.5	77.1
N24News	36.9	50.0	73.1	73.4	77.6	42.5	42.3	45.6	77.9	43.3	76.5	80.2
HatefulMemes	48.1	56.8	57.3	67.5	72.3	52.5	51.2	66.4	69.0	51.2	71.2	75.7
VOC2007	53.0	66.4	71.8	87.5	88.2	75.1	68.3	54.4	83.3	77.2	87.9	89.3
SUN397	54.8	73.8	71.2	78.0	76.1	46.1	61.6	73.1	75.0	62.7	76.6	77.2
Place365	31.6	41.0	40.4	43.4	44.9	16.2	37.1	38.6	44.7	33.7	44.9	47.1
ImageNet-A	41.6	45.4	43.3	53.4	51.8	29.9	52.4	53.0	60.4	24.5	57.6	58.0
ImageNet-R	82.8	72.0	85.6	85.1	82.4	62.7	87.7	83.0	86.7	68.0	86.5	84.6
ObjectNet	46.2	73.2	65.5	55.5	55.7	37.8	62.9	75.4	68.2	60.9	52.3	53.3
Country211	23.6	29.3	26.0	29.8	29.8	25.7	32.5	33.5	30.8	20.7	30.8	32.3
Image QA (I-QA)												
OK-VQA	17.4	57.8	62.7	65.2	66.8	24.2	29.3	62.0	70.4	32.7	70.5	70.4
A-OKVQA	12.5	48.1	52.2	55.4	53.6	15.6	20.0	53.9	59.6	25.0	57.8	59.3
DocVQA	17.5	84.1	81.6	91.3	92.6	22.8	25.8	89.6	93.3	18.6	93.4	92.3
InfographicsVQA	8.5	58.6	49.9	68.2	63.5	11.7	12.4	66.6	68.9	17.0	74.7	72.9
ChartQA	13.3	40.1	47.7	58.7	58.5	12.4	14.0	46.9	65.7	12.5	64.5	63.1
Visual7W	7.4	48.5	55.2	49.5	48.1	4.8	11.9	47.9	63.7	18.7	53.3	50.8
ScienceQA	25.3	44.4	48.7	43.2	43.9	25.5	33.2	52.2	56.9	21.8	54.4	52.6
VizWiz	34.8	54.6	52.5	52.3	51.0	26.0	41.8	54.1	55.9	31.2	54.8	52.1
GQA	28.5	65.4	79.1	67.4	67.9	37.1	48.8	69.7	78.4	50.2	70.1	70.0
TextVQA	36.1	80.6	83.0	79.5	81.2	44.9	58.5	84.3	85.9	29.9	84.0	84.0
Image-Text Retrieval (I-RET)												
VisDial	52.8	50.5	77.8	71.0	73.9	50.6	56.2	54.7	80.8	19.7	74.8	76.1
CIRR	13.2	42.9	41.1	47.8	48.8	14.4	16.2	47.1	49.3	31.4	48.0	52.5
VisualNews_t2i	60.1	54.9	70.1	70.7	72.6	37.7	63.3	61.3	74.5	45.3	74.3	75.2
VisualNews_i2t	58.4	59.4	76.4	73.4	75.1	43.8	64.2	61.1	82.3	36.3	75.7	77.3
MSCOCO_t2i	59.9	59.2	74.3	65.6	67.1	51.2	63.6	65.2	76.2	63.5	70.7	70.7
MSCOCO_i2t	56.2	59.7	71.7	67.2	67.4	40.7	59.0	62.0	74.1	55.1	70.1	69.7
NIGHTS	64.6	56.7	66.3	68.5	67.3	59.2	63.7	61.3	64.8	58.2	67.2	67.6
WebQA	90.2	84.7	90.5	89.1	88.9	77.1	90.9	84.0	88.6	51.5	90.3	90.1
FashionIQ	8.6	14.9	16.5	19.9	19.1	8.3	11.4	19.3	19.4	6.5	18.8	22.9
Wiki-SS-NQ	87.0	52.6	74.2	69.1	62.3	84.2	86.6	55.0	83.3	51.1	64.4	63.6
OVEN	72.7	75.2	75.8	74.0	74.1	65.2	74.6	77.2	79.1	57.2	77.3	76.9
EDIS	81.2	53.8	85.3	79.9	79.2	61.7	83.0	58.4	90.0	53.1	88.3	86.9
Visual Grounding (VG)												
MSCOCO	35.1	45.5	50.9	67.9	68.0	40.0	42.4	44.2	67.2	25.1	74.1	76.1
RefCOCO	55.1	68.8	76.6	75.3	84.8	67.8	65.9	73.5	83.7	49.7	87.3	91.0
RefCOCO-Matching	65.3	75.7	79.2	80.3	85.6	74.6	69.8	82.6	91.6	83.1	91.6	90.6
Visual7W-Pointing	42.7	55.8	69.6	67.8	72.9	59.6	62.0	60.2	81.1	49.7	79.6	76.6

Table 16: Individual performances on image modality tasks. The method order is identical to that in Tab. 1.

	Nemotron (3B)	LCO-Emb (3B)	e5-omni (3B)	Uni-Omni (3B)	Syn-Omni (3B)	OmniEmbed (7B)	Multivent (7B)	LCO Emb (7B)	e5-omni (7B)	WAVE (7B)	Uni-Omni (7B)	Syn-Omni (7B)
Average												
Overall (23)	36.5	43.7	40.6	40.4	40.5	35.0	40.2	45.5	44.3	39.8	40.2	41.4
V-CLS (5)	43.1	44.6	37.6	47.6	44.7	36.4	50.7	47.6	49.4	48.9	49.5	47.0
T2V RET (5)	34.8	34.2	38.9	27.9	29.3	33.7	35.9	36.9	36.1	30.0	25.0	28.5
V2T RET (5)	32.0	34.3	33.2	38.3	39.2	27.9	34.7	35.9	40.8	35.1	40.5	41.6
M-RET (3)	25.6	47.2	41.0	36.9	37.5	28.7	28.1	47.1	34.1	40.1	32.4	36.1
V-QA (5)	46.9	58.1	52.0	51.1	51.7	48.3	51.8	60.2	61.1	44.9	53.5	53.8
Video Classification (V-CLS)												
K700	38.5	42.6	42.4	46.3	43.9	31.5	53.3	43.2	50.1	51.4	52.0	48.0
SmthSmthV2	40.0	44.1	35.2	45.9	45.9	34.6	54.8	46.5	46.1	47.2	48.8	51.1
HMDB51	48.0	53.2	34.1	52.1	47.3	26.2	53.3	57.7	56.3	53.4	54.7	44.2
UCF101	62.7	64.2	51.9	71.5	63.1	62.2	68.2	67.8	69.5	72.0	66.6	68.3
Breakfast	26.3	18.7	24.5	22.2	23.1	27.3	24.0	22.6	24.9	20.3	25.4	23.6
Text → Video Retrieval (T2V RET)												
DiDeMo	36.0	33.1	41.6	27.4	27.4	27.6	34.7	34.7	35.7	25.6	23.1	27.3
MSR-VTT	37.8	37.4	40.0	39.6	43.1	40.2	40.5	41.2	42.1	34.6	37.6	43.9
MSVD	55.8	58.4	62.1	43.3	46.7	56.4	56.9	60.0	56.7	50.9	39.9	45.1
VATEX	28.8	29.1	34.0	25.5	24.3	28.1	30.1	32.3	30.8	29.4	23.0	21.3
YouCook2	15.5	13.0	16.8	3.7	5.0	16.0	17.4	16.1	15.3	9.7	1.7	4.7
Video → Retrieval (V2T RET)												
DiDeMo	32.0	32.4	31.9	42.5	42.0	27.7	35.7	33.6	43.0	32.7	44.3	46.4
MSR-VTT	34.3	38.8	37.6	42.4	45.8	33.2	37.3	37.9	44.2	39.0	47.0	48.1
MSVD	53.6	56.9	57.6	60.7	61.5	49.7	56.9	60.1	63.1	58.4	62.5	63.7
VATEX	25.6	28.3	28.4	29.9	31.4	21.5	28.5	30.0	37.2	30.8	31.6	32.0
YouCook2	14.4	15.4	10.6	15.8	15.3	7.4	15.1	17.6	16.4	14.7	17.3	18.0
Moment Retrieval (M-RET)												
QVHighlight	26.9	70.3	60.4	51.9	53.4	29.9	30.0	69.2	44.1	56.0	41.5	50.9
Charades-STA	11.6	25.2	19.5	18.0	16.6	14.3	13.1	24.3	16.9	26.4	16.8	14.9
MomentSeeker	38.3	46.1	43.1	40.8	42.6	41.9	41.2	47.8	41.4	38.0	39.0	42.7
Video QA (V-QA)												
MVBench	42.3	58.0	51.5	50.9	53.1	43.6	47.2	60.9	60.1	39.3	54.7	54.5
Video-MME	40.1	46.1	43.0	41.3	41.6	38.4	40.9	49.7	49.0	32.1	43.6	44.0
NExTQA	43.9	67.4	58.9	55.5	59.4	45.9	48.1	70.3	71.1	38.4	60.8	60.0
EgoSchema	46.6	43.8	45.6	34.8	30.6	46.4	51.4	42.8	51.4	48.0	34.2	35.4
ActivityNetQA	61.6	75.1	61.1	72.9	73.9	67.3	71.5	77.3	74.1	66.5	74.2	75.1

Table 17: Individual performances on video modality tasks. The method order is identical to that in Tab. 1.

	Nemotron (3B)	LCO-Emb (3B)	e5-omni (3B)	Uni-Omni (3B)	Syn-Omni (3B)	OmniEmbed (7B)	Multivent (7B)	LCO Emb (7B)	e5-omni (7B)	WAVE (7B)	Uni-Omni (7B)	Syn-Omni (7B)
Average												
Overall (12)	24.5	42.1	37.4	43.1	45.6	31.6	39.5	45.2	44.3	33.4	47.5	48.7
A-CLS (3)	40.7	67.5	51.5	59.5	62.0	36.0	52.9	71.1	62.8	56.3	67.4	67.2
T2A RET (4)	6.4	19.7	26.0	28.7	30.8	25.5	28.6	24.5	30.2	22.5	32.5	34.9
A2T RET (3)	6.8	17.3	14.2	30.0	32.4	17.2	21.0	18.7	20.7	17.7	32.5	34.0
A-QA (2)	44.0	63.8	57.9	54.4	57.0	47.8	55.4	66.4	63.7	37.0	57.6	58.5
Audio Classification (A-CLS)												
ESC50	73.6	77.8	78.2	86.8	87.0	51.5	77.8	78.8	81.9	82.8	88.0	88.5
GTZAN	28.6	74.8	44.1	52.1	57.9	34.5	51.4	82.8	61.4	57.9	66.9	63.4
NSynth	20.0	50.0	32.1	39.6	41.0	22.0	29.5	51.6	45.0	28.1	47.3	49.7
Text → Audio Retrieval (T2A RET)												
AudioCaps	9.4	23.3	33.9	39.9	43.2	39.0	33.5	26.9	38.6	32.1	41.2	44.9
Clotho	7.2	17.2	21.6	20.4	20.6	21.8	21.5	19.8	22.7	19.1	21.3	21.8
MusicCaps	3.5	8.0	10.5	19.8	23.4	11.8	13.2	11.9	14.8	8.7	20.6	22.6
SoundDescs	5.3	30.4	38.1	34.6	36.1	29.4	46.2	39.3	44.5	29.9	46.9	50.5
Audio → Text Retrieval (A2T RET)												
AudioCaps	10.6	26.0	26.0	49.4	52.8	33.7	34.8	30.0	36.8	26.6	54.2	57.6
Clotho	7.4	17.4	13.3	23.4	23.6	15.2	19.4	18.0	17.1	19.1	26.9	26.5
MusicCaps	2.3	8.3	3.3	17.2	20.7	2.7	8.9	8.2	8.0	7.3	16.4	18.0
Audio QA (A-QA)												
MMAU	50.7	70.5	63.9	59.9	62.1	50.0	58.7	74.9	70.5	36.1	64.8	61.8
MMAR	37.4	57.1	52.0	48.8	52.0	45.5	52.1	58.0	56.9	38.0	50.5	55.3

Table 18: Individual performances on audio modality tasks. The method order is identical to that in Tab. 1.

	Nemotron (3B)	LCO-Emb (3B)	e5-omni (3B)	Uni-Omni (3B)	Syn-Omni (3B)	OmniEmbed (7B)	Multivent (7B)	LCO Emb (7B)	e5-omni (7B)	WAVE (7B)	Uni-Omni (7B)	Syn-Omni (7B)
Average												
Overall (10)	28.5	35.5	31.6	40.4	43.4	29.0	34.4	35.5	42.3	38.1	43.9	44.6
A2V RET (2)	5.2	11.1	7.0	17.4	19.1	5.2	11.4	12.6	7.2	8.8	22.0	23.1
V2A RET (2)	7.3	10.5	9.2	14.3	17.9	13.5	16.1	13.8	14.8	19.9	20.1	18.9
T2VA RET (2)	46.9	51.1	64.4	63.3	64.2	57.2	48.9	51.7	65.7	52.6	64.7	64.9
VA2T RET (2)	41.4	46.6	30.5	62.5	63.0	27.3	49.1	46.3	62.4	56.9	64.3	63.9
AV-QA (2)	41.5	58.4	46.9	44.8	52.6	41.8	46.7	53.3	61.6	52.4	48.5	52.0
Audio → Video Retrieval (A2V RET)												
AVE	6.7	12.9	10.2	18.2	19.4	8.2	11.2	15.2	9.5	10.0	23.4	23.9
VALOR32k	3.7	9.2	3.9	16.6	18.8	2.2	11.6	9.9	5.0	7.7	20.6	22.2
Video → Audio Retrieval (V2A RET)												
AVE	10.9	12.7	10.9	12.9	19.9	16.4	17.4	14.7	15.9	22.1	20.6	15.9
VALOR32k	3.7	8.4	7.5	15.6	16.0	10.5	14.8	13.0	13.6	17.7	19.6	21.9
Text → Audiovisual Retrieval (T2VA RET)												
AVHBench	56.2	60.6	75.0	74.9	75.7	67.3	57.6	61.1	76.5	62.4	75.4	76.0
VALOR32k	37.6	41.6	53.8	51.6	52.7	47.1	40.1	42.4	54.9	42.7	54.0	53.9
Audiovisual → Text Retrieval (AV2T RET)												
AVHBench	50.2	53.2	23.6	73.5	74.8	36.6	58.6	54.6	71.1	67.1	74.8	75.4
VALOR32k	32.6	39.9	37.4	51.6	51.3	18.1	39.6	38.0	53.8	46.8	53.8	52.4
Audiovisual QA (AV-QA)												
AVHBenchQA	53.4	69.8	55.8	63.8	65.5	52.2	60.7	60.5	79.5	76.2	60.2	66.7
DailyOmni	29.5	47.0	38.1	25.8	39.7	31.4	32.6	46.2	43.8	28.6	36.8	37.2

Table 19: Individual performances on audiovisual modality tasks. The method order is identical to that in Tab. 1.