AVEMBEDBENCH: A BENCHMARK FOR AUDIO-VISUAL JOINT RETRIEVAL IN OMNI-EMBEDDING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent omni-embedding models support both visual and audio inputs, yet whether they effectively integrate both modalities for retrieval remains unclear. Existing audio-visual benchmarks primarily evaluate generative models through question answering. However, their reliance on textual responses and focus on answer accuracy rather than retrieval rankings limit their applicability to evaluating joint retrieval in embedding models. To address these limitations, we introduce AVEmbedBench, to our knowledge the first benchmark specifically designed to evaluate audio–visual joint retrieval in omni-embedding models. In AVEmbedBench, Each query combines visual and audio-semantic constraints that the target video must jointly satisfy. To discourage unimodal shortcuts, we select clips with weakly correlated visual and audio semantics and pair each target with two hard negatives, one matching only the visual constraint and the other only the audio constraint. To construct the benchmark at scale, we develop an automated pipeline for audio-visual decoupled video filtering, joint query generation, and modality-specific hard negative mining. Evaluations of recent omni-embedding models reveal a consistent visual bias: models favor visual similarity despite audio mismatch, underscoring the gap between multimodal input support and effective joint retrieval, and motivating more balanced audio–visual integration.

1 INTRODUCTION

The growth of multimodal content has made retrieval across text, images, audio, and video increasingly important for multimedia information access Wang et al. (2024a). Multimodal embedding models map heterogeneous inputs into a shared semantic space for similarity-based retrieval Jia et al. (2021); Wang et al. (2023). Beyond image–text alignment, this paradigm now encompasses audio, video, and their combinations Shvetsova et al. (2022); Wang et al. (2023). Audio enriches video representations with complementary cues, including spoken content and environmental sounds Rouditchenko et al. (2021). **However, supporting both visual and audio inputs does not guarantee effective integration.** When queries impose both visual and audio constraints, reliance on either modality alone may favor partial matches over jointly relevant candidates. Evaluating visual–audio joint retrieval is therefore essential to assessing whether these models capture joint semantic relevance rather than single-modality similarity.

Recent benchmarks have been proposed to assess joint audio-visual understanding in large language models (Kim et al., 2025; Leng et al., 2025; Zhang et al., 2026). A representative evaluation strategy uses videos with differing auditory and visual evidence and asks modality-specific questions to probe whether models hallucinate sounds from visual cues or visual events from auditory cues. For example, as illustrated in Figure 1 (a), a video shows a dog while an off-screen bird chirps and the dog remains silent. Answering "yes" to "Is the dog making sound in the audio?" constitutes video-driven audio hallucination; conversely, answering "yes" to "Is the chirping bird visible in the video?" constitutes audio-driven video hallucination (Kim et al., 2025). Subsequent studies report persistent cross-modal hallucinations on these benchmarks(Chen et al., 2026b; Chung et al., 2026; Chaubey et al., 2026), highlighting over-reliance on unimodal priors and spurious cross-modal associations.

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

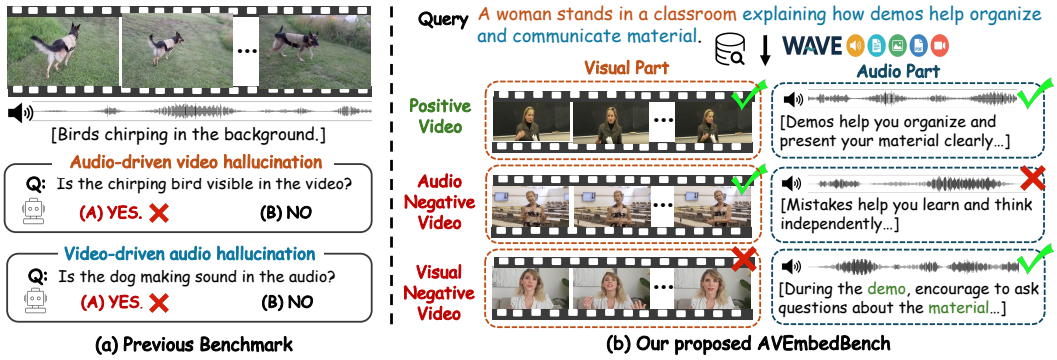


Figure 1: **Differences between our proposed AVEmbedBench and existing benchmarks.** Existing benchmarks assess audio-visual understanding through question answering. In contrast, AVEmbedBench evaluates retrieval under joint visual and speech-semantic constraints. By pairing each query with a fully matched video and hard negatives matching only one modality, it probes joint audio-visual retrieval and unimodal reliance in omni-embedding models.

Despite these advancements, a critical limitation remains: AVHBench is not directly applicable to omni-embedding models due to two fundamental mismatches. **First, AVHBench requires Yes/No or free-form textual answers, whereas embedding models produce continuous vector representations.** Converting such outputs into similarity scores against "Yes" and "No" embeddings may introduce sensitivity to answer phrasing and prompt design. **Second, AVHBench evaluates individual question-answering decisions rather than the ability to rank target videos among candidates.** Therefore, detecting audio–visual inconsistency does not necessarily indicate that embeddings can distinguish fully matched videos from distractors matching only one modality, making QA accuracy only an indirect measure of joint audio–visual understanding in retrieval settings.

To address this deficiency, in this work, we introduce AVEmbedBench, to our knowledge the first benchmark specifically designed to evaluate audio-visual joint retrieval in omni-embedding models. In AVEmbedBench, **each query imposes both visual and speech-semantic constraints, requiring the model to retrieve the video that satisfies them jointly.** For example, as shown in Figure 1 (b), the query "A woman stands in a classroom explaining how demos help organize and communicate material" combines a visual description with spoken semantic content. To discourage reliance on a single modality, AVEmbedBench selects video clips with weakly correlated visual and speech semantics, such that neither modality can be reliably inferred from the other. For each query, we further construct hard negatives that match only one modality: either the visual content or the speech semantics. To build this benchmark, we develop an automated pipeline that **selects videos with semantically decoupled visual and speech content**, **generates joint queries using Qwen3.6-27B** (Qwen Team, 2026), and **mines modality-specific hard negatives**.

We evaluate recent Omni-Embedding models that support joint visual and audio inputs on our proposed AVEmbedBench, including the open-source models Omni-Embed-Nemotron Xu et al. (2025a), LCO-Embedding-Omni Xiao et al. (2025a), E5-Omni Chen et al. (2026a), WAVE Tang et al. (2026), and Jina-Omni Hönicke et al. (2026a), as well as the commercial model Gemini Embedding 2 Shanbhogue et al. (2026a). Experiments show that these models **exhibit some capability for visual-audio joint retrieval but remain susceptible to hard negatives matching only one modality**, struggling to consistently distinguish single-modality similarity from joint relevance. Further analysis reveals a consistent **visual bias** across the evaluated models: **visually matching negatives are more likely to outrank the positive despite mismatched audio.** Modality ablations further suggest that this **bias stems not only from limited audio-semantic understanding but also from interference introduced by joint visual–audio encoding**, making otherwise retrievable audio-semantic information less accessible. Our key contributions are as follows:

- We introduce AVEmbedBench, to our knowledge the first benchmark specifically designed to evaluate visual–audio joint retrieval in Omni-Embedding models. It comprises 275,409 retrieval instances spanning 637,481 unique video clips, challenging models to distinguish jointly relevant videos from hard negatives matching only one modality.

- We develop an automated data construction pipeline integrating audio-visual decoupled video filtering, joint query generation, and modality-specific hard negative mining, enabling scalable benchmark construction.
- We evaluate recent Omni-Embedding models on AVEmbedBench and systematically analyze their retrieval performance. Our findings reveal a consistent visual bias: models favor visual similarity despite audio mismatch. Modality ablations suggest that this bias reflects not only limitations in audio-semantic understanding but also interference introduced by joint visual–audio encoding.

2 Related Work

2.1 Multimodal Embedding Models.

Multimodal embedding models map heterogeneous inputs into a shared semantic space for cross-modal retrieval. For example, CLIP (Radford et al., 2021) aligns images and text through contrastive learning, while CLIP4Clip (Luo et al., 2022) extends this paradigm to text-video retrieval. Recent omni-embedding models further incorporate audio and multimodal combinations, advancing toward any-to-any retrieval. For instance, WAVE-7B (Tang et al., 2026), Omni-Embed-Nemotron-3B (Xu et al., 2025b), and LCO-Embedding-Omni-3B/7B (Xiao et al., 2025b) adapt pretrained multimodal backbones through contrastive learning. Gemini Embedding 2 (Shanbhogue et al., 2026b) supports native interleaved multimodal encoding, whereas jina-embeddings-v5-omni (Hönicke et al., 2026b) connects frozen modality encoders through lightweight projectors. Complementing these approaches, e5-omni (Chen et al., 2026a) explicitly calibrates cross-modal alignment. By incorporating audio alongside visual content, these models can potentially capture complementary evidence for richer retrieval representations. **We therefore focus on omni-embedding models and evaluate whether their embeddings effectively integrate auditory and visual information.**

2.2 Multimodal Embedding Benchmarks.

Existing benchmarks evaluate multimodal embeddings across diverse tasks and modalities. For example, MMEB (Jiang et al., 2025) unifies image-text classification, question answering, retrieval, and visual grounding within a retrieval-based framework, while MMEB-V2 (Meng et al., 2025) extends this coverage to videos and visual documents. MIEB (Xiao et al., 2025c) further provides a broad evaluation of image and image-text representations, including clustering, retrieval, and compositional understanding. On the audio side, AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020) are widely used to assess audio-text retrieval. **Although these benchmarks measure complementary aspects of multimodal representation quality, they do not directly assess whether embeddings jointly capture auditory and visual evidence.** Our benchmark complements these evaluations by specifically targeting joint audio-visual understanding in omni-embedding models through retrieval.

2.3 Audio-Visual Joint Understanding Benchmarks.

Several benchmarks have been proposed to assess joint audio-visual understanding in large language models. For example, AVHBench (Kim et al., 2025) evaluates cross-modal hallucinations by examining whether models infer nonexistent sounds from visual cues or hallucinate visual events from auditory cues. CMM (Leng et al., 2025) systematically investigates hallucinations across language, vision, and audio, highlighting over-reliance on unimodal priors and spurious inter-modal correlations. More recently, SVHalluc (Zhang et al., 2026) focuses on speech-vision hallucinations, assessing whether models correctly establish semantic and temporal correspondence between speech content and visual evidence. **However, these benchmarks primarily evaluate generative audio-visual models through question-answering tasks that require textual responses, emphasizing yes/no predictions rather than retrieval rankings.** In contrast, we focuses on omni-embedding models, which produce vector representations rather than natural-language responses. We assess their joint audio-visual understanding in a retrieval setting, examining whether they can rank videos satisfying both visual and audio constraints above distractors matching only one modality.

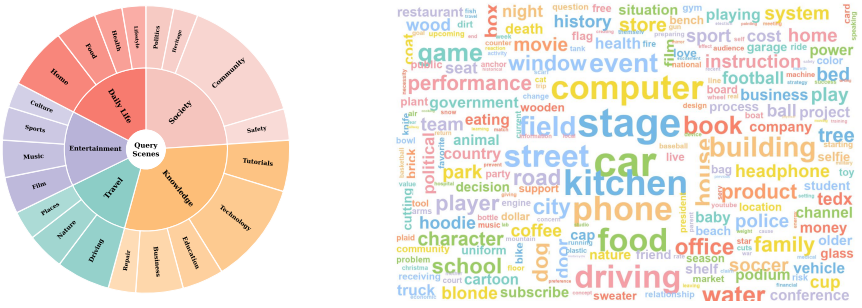


Figure 2: **Data diversity of AVEmbedBench.** (Left) Hierarchical distribution of query scenes across broad categories and fine-grained subcategories. (Right) Query word cloud summarizing frequent scene, object, activity, and topic terms, with word size proportional to frequency.

Table 1: **Dataset statistics of AVEmbedBench**, comprising large-scale video triplets paired with queries describing both visual content and speech semantics.

Category	Statistic	Value
Video	Triplet instances	275,409
	Unique clips	637,481
	Unique source videos	396,916
	Average clip duration (seconds)	7.48
Query	Average visual-part word count	10.45
	Average speech-part word count	11.23

3 AVEmbedBench: Audio-visual Joint Retrieval Benchmark

3.1 Overview of AVEmbedBench

We introduce AVEmbedBench, to our knowledge the first benchmark specifically designed to evaluate audio-visual joint retrieval in omni-embedding models. Unlike prior benchmarks that assess joint understanding through textual answers to audio-visual questions, AVEmbedBench evaluates retrieval ranking using queries that impose both visual and speech-semantic constraints. To discourage unimodal reliance, we select video clips with weakly correlated visual and speech semantics and construct two types of hard negatives for each query: visually matched but speech-mismatched videos, and speech-matched but visually mismatched videos. Models must therefore distinguish targets satisfying both constraints from single-modality distractors. Figure 2 illustrates the diversity of AVEmbedBench across virous scene categories. Dataset statistics are summarized in Table 1. Overall, AVEmbedBench comprises **275,409 retrieval instances** spanning **637,481 unique video clips**, each pairing a joint query with one positive video and two modality-specific hard negatives.

3.2 Data Construction Pipeline

We construct AVEmbedBench through a three-stage automated pipeline: semantic similarity filtering, joint query generation, and hard negative video mining.

Stage 1: Audio-visual Decoupled Video Filtering. We begin by collecting videos from **InternVideo2_Vid_Text** Wang et al. (2024b), a large-scale dataset comprising approximately 61 million clips from 2.12 million videos, annotated with visual captions, speech transcripts and video URLs. Since its annotations are automat ically generated, the dataset inevitably contains substantial noise, including corrupted captions and uninformative speech transcripts. We therefore apply rule-based filtering to remove clearly unsuitable clips.

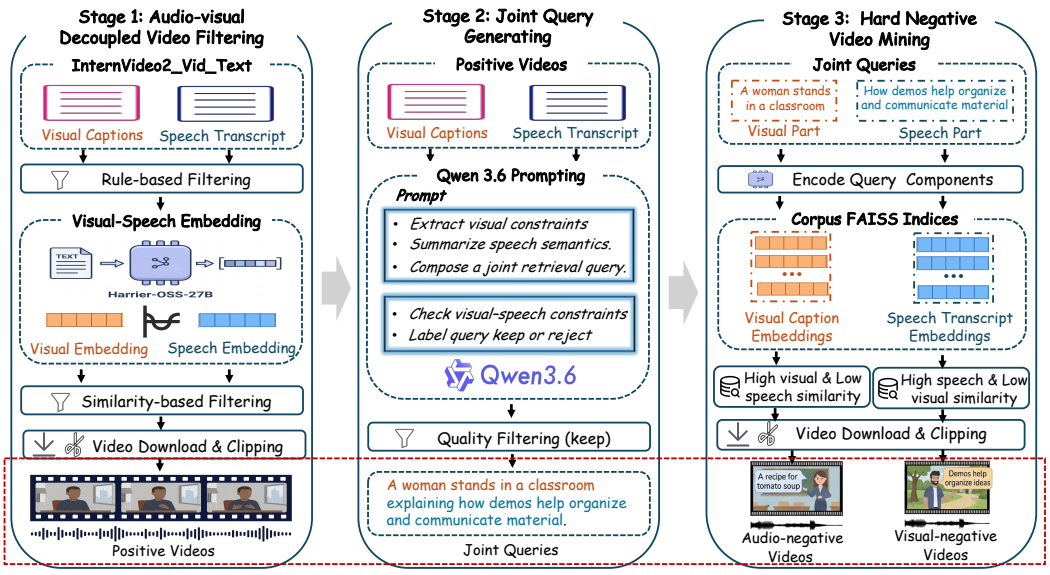


Figure 3: **Dataset construction pipeline.** We construct AVEmbedBench via three automated stages. Stage 1 selects videos with weakly correlated visual and speech content through rule-based and similarity-based filtering; Stage 2 generates and validates joint queries integrating visual constraints with speech semantics; Stage 3 mines hard negatives matching one modality but not the other to assess reliance on unimodal shortcuts.

A key challenge in constructing our dataset is to identify videos in which the visual and speech contents are semantically decoupled. When the two modalities are highly correlated, models may infer one modality from the other without genuinely performing joint audio-visual reasoning. Given the scale of the candidate pool, manual assessment of such cross-modal relationships is impractical. We therefore use Harrier-OSS-27B Microsoft (2026) to separately encode each clip's visual caption and speech transcript, and compute their cosine similarity as a proxy for semantic correlation. A higher similarity score indicates stronger semantic alignment between the visual scene and speech content, whereas a lower score suggests greater cross-modal decoupling.

After filtering clips based on audio-visual semantic similarity, we download the corresponding YouTube videos using the URLs provided in the source annotations. This filtering process retains approximately 490,000 video clips that meet the selection criteria.

Stage 2: Joint Query Generating. After selecting videos with weakly correlated visual and speech semantics, we construct queries that impose constraints on both modalities. Although InternVideo2_Vid_Text Wang et al. (2024b) provides visual captions and speech transcripts, these annotations are independent, and directly concatenating them may yield unnatural descriptions unsuitable for retrieval. We therefore prompt Qwen3.6-27B (Qwen Team, 2026) to extract visual conditions from the captions, summarize the spoken content, and integrate both into coherent retrieval queries. Queries follow the structure "[visual description] while [speech description]," preserving explicit constraints on each modality.

Qwen3.6-27B also assigns each sample a quality label: `keep` or `reject`. Of approximately 490,000 generated queries, 275,409 are labeled `keep` and retained, yielding audio-visual joint query–clip pairs for retrieval evaluation.

Stage 3: Hard Negative Video Mining. To assess audio-visual joint understanding and diagnose unimodal reliance, we mine two types of hard-negative videos for each query, each matching one modality while differing in the other. Specifically, *audio-negative* videos match the query's visual constraints but contain semantically unrelated speech, probing reliance on visual shortcuts. Conversely, *visual-negative* videos match the query's speech-semantic constraints but contain unrelated visual content, probing reliance on audio shortcuts. To enable efficient hard-negative mining across

the full video corpus, we encode the visual captions and speech transcripts in the source annotations using Harrier-OSS-27B Microsoft (2026) and build separate FAISS indices for the resulting caption and transcript embeddings.

Audio-negative video mining. These negatives retain query-relevant visual content but differ in speech semantics. We encode the visual component of each query using Harrier-OSS-27B Microsoft (2026) and retrieve the top-5 most similar candidates from the visual index. We then rank these candidates by their speech-semantic similarity to the positive video and select those with the lowest similarity. These negatives are designed to challenge visual-only matching and test whether models use speech semantics to identify the target.

Visual-negative video mining. Conversely, these negatives match the query's speech semantics but differ in visual content. We embed the speech component of each query and retrieve the top-5 most similar candidates from the speech index. Among these, we select candidates with the lowest visual-semantic similarity to the positive video. These negatives test whether models use visual evidence rather than relying solely on speech. For both types, we filter unavailable source videos and extract the selected clips to assemble retrieval instances.

4 EXPERIMENTS

4.1 EXPERIMENT SETUP

Models. AVHEmbedBench is evaluated on six different recent multimodal embedding models, mostly referred to in current literature, and capable of embedding both audio and visual signals simultaneously. The models are Omni-Embed-Nemotron Xu et al. (2025a), LCO-Embedding-Omni Xiao et al. (2025a), E5-Omni Chen et al. (2026a), WAVE Tang et al. (2026), Jina-Omni Hönicke et al. (2026a), and Gemini Embedding 2 Shanbhogue et al. (2026a). For the five open-source models, we encode text queries and candidate videos separately, following each model's input conventions to process sampled frames and the corresponding audio. For Gemini Embedding 2, we obtain embeddings through its API by submitting text queries and videos with audio separately. All models are evaluated in a zero-shot setting, without task-specific training or fine-tuning.

Evaluation metrics. For each query, we rank candidate videos by the similarity between their embeddings and the query embedding. We evaluate overall retrieval performance using Recall@K (R@K) for $K \in \{1, 5, 10\}$ and Mean Reciprocal Rank (MRR). R@K measures the proportion of queries for which the positive video appears among the top-K retrieved candidates, while MRR is the average reciprocal rank of the positive video across all queries.

Moreover, we assess modality-related retrieval bias using three complementary metrics: **Joint Ranking Accuracy (JRA), Visual Negative Outranking Rate ($\mathbf{OR}_v$), and Audio Negative Outranking Rate ($\mathbf{OR}_a$).** Let s_i^+, s_i^v, and s_i^a denote the scores of the positive video, visual-only matching negative, and speech-only matching negative, respectively. For N queries:

$$\mathrm{JRA} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[s_i^+ > \max(s_i^v, s_i^a)], \quad \mathrm{OR}_v = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[s_i^v > s_i^+], \quad \mathrm{OR}_a = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[s_i^a > s_i^+], \tag{1}$$

where $\mathbb{I}[\cdot]$ is the indicator function. JRA measures whether the positive outranks both negative types, assessing retrieval under joint visual and speech-semantic constraints, whereas OR_v and OR_a capture failures to reject speech-semantic and visual mismatches, respectively. Higher JRA and lower outranking rates indicate better performance.

4.2 RESULTS AND DISCUSSIONS

4.2.1 MAIN RESULTS

We evaluate existing Omni-Embedding models on our proposed AVEmbedBench for visual–audio joint retrieval, where relevant videos must satisfy both visual and audio query constraints. Table 2 reports retrieval performance with and without modality-specific hard negatives. In the setting without hard negatives, we remove only the two hard negatives associated with each query, leaving all other candidates unchanged.

Table 2: **Overall retrieval performance of Omni-Embedding models on AVEmbedBench with and without hard negatives.** "Without Hard Negatives" excludes only the two hard negatives associated with each query, leaving all other can didates unchanged. ↑ indicates higher is better.

Model	With Hard Negatives				Without Hard Negatives			
	R@1 ↑	R@5 ↑	R@10 ↑	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	MRR ↑
Omni-Embed-Nemotron	**86.70**	92.87	94.21	91.09	**90.84**	94.07	95.01	91.77
LCO-Embedding-Omni	**55.43**	57.52	63.46	61.61	**59.12**	62.47	68.82	65.84
E5-Omni	**82.27**	91.26	93.29	86.36	**84.56**	92.65	94.43	88.25
WAVE	**63.48**	79.73	84.23	70.90	**67.28**	82.71	86.77	74.31
Jina-Omni	**52.74**	63.93	69.62	58.51	**56.27**	69.07	75.30	62.76
Gemini Embedding 2	**89.47**	93.92	94.50	91.40	**91.98**	94.77	95.26	93.34

Table 3: **Modality bias evaluation of Omni-Embedding models on AVEmbedBench.** ↑ and ↓ indicate that higher and lower values are better, respectively.

Model	Joint Ranking	Outranking Rate		Top-1 Selection	
	JRA ↑	OR_v ↓	OR_a ↓	$\mathrm{NSR}_v@1$ ↓	$\mathrm{NSR}_a@1$ ↓
Omni-Embed-Nemotron	88.82	7.27	4.39	7.12	3.98
LCO-Embedding-Omni	67.38	19.14	9.88	17.39	8.14
E5-Omni	84.27	13.85	2.21	13.69	1.95
WAVE	75.80	20.32	4.61	19.95	3.90
Jina-Omni	62.79	20.21	14.04	19.06	12.37
Gemini Embedding 2	89.74	6.45	3.11	6.30	2.88

Performance varies considerably across models. Gemini Embedding 2 achieves the best overall performance, but its R@1 remains at 89.47% with hard negatives, highlighting the difficulty of consistently ranking jointly relevant videos first. Removing hard negatives improves performance for all models; for instance, the R@1 of Omni-Embed-Nemotron increases from 86.70% to 90.84%. This improvement indicates that candidates satisfying only one modality constraint can interfere with the ranking of jointly relevant videos. Overall, current Omni-Embedding models demonstrate some capability for visual–audio joint retrieval but remain susceptible to single-modality distractors. These results highlight the challenge of distinguishing single-modality matches from candidates satisfying both visual and audio constraints.

4.2.2 Modality Basis Evaluation

We further investigate modality bias in visual–audio joint retrieval. Here, modality bias refers to a model's tendency to favor one modality over the other when ranking retrieved candidates. Specifically, we evaluate models on our proposed AVEmbedBench using three complementary groups of metrics. Joint Ranking Accuracy (JRA) measures the proportion of queries for which the positive video outranks both associated hard negatives. Outranking Rates (OR_v and OR_a) measure how often visual-only and audio-only matching negatives, respectively, outrank the positive. Top-1 Negative Selection Rates ($\mathrm{NSR}_v@1$ and $\mathrm{NSR}_a@1$) measure how often these negatives occupy the top rank in the full candidate pool. Together, these metrics characterize joint ranking accuracy and modality-specific retrieval errors.

As shown in Table 3, Omni-Embed-Nemotron achieves the highest JRA of 88.82%, exceeding E5-Omni by 4.55 percentage points and demonstrating a stronger ability to distinguish jointly relevant videos from single-modality matches. Despite performance differences, all evaluated models exhibit a consistent asymmetry: OR_v exceeds OR_a, indicating greater susceptibility to negatives that satisfy the visual constraint but violate the audio constraint. This asymmetry also appears in top-1 selection, where $\mathrm{NSR}_v@1$ is consistently higher than $\mathrm{NSR}_a@1$, showing that visual-only matches

Table 4: **Modality basis analysis of Omni-Embedding models on AVEmbedBench.** We compare audio-only and joint audio–visual embeddings using audio-related query clauses, with query-associated visual negatives excluded. ↑ indicates that higher values are better.

Model	Audio				Visual + Audio			
	R@1 ↑	R@5 ↑	R@10 ↑	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	MRR ↑
Omni-Embed-Nemotron	**66.32**	79.35	83.62	72.84	**59.85**	75.73	80.96	69.18
LCO-Embedding-Omni	**48.40**	56.35	60.26	52.04	**44.71**	47.88	49.45	46.39
E5-Omni	**67.10**	80.27	84.01	73.16	**57.30**	72.11	76.81	64.19
WAVE	**56.11**	68.60	74.57	62.40	**53.53**	66.13	72.88	60.08
Jina-Omni	**42.73**	54.34	59.31	48.48	**32.92**	35.12	36.30	34.14
Gemini Embedding 2	**76.67**	83.94	85.42	79.96	**68.14**	78.65	81.27	72.98

disproportionately affect the final retrieval outcome. Overall, these results suggest a **visual bias problem: models tend to favor visual similarity despite audio mismatch**, limiting their ability to distinguish single-modality matches from videos satisfying both visual and audio constraints.

4.2.3 Modality Basis Analysis

We next investigate whether visual bias in audio–visual joint retrieval stems from limited speech-semantic representations or interference introduced by joint encoding. Specifically, we extract the audio-related clause from each query and retrieve candidate videos using either audio-only or joint audio–visual embeddings. For each query, we exclude its associated visual-negative candidates, as these remain valid matches for the audio-only query. As shown in Table 4, audio-only retrieval consistently outperforms joint retrieval across the evaluated models. For instance, E5-Omni achieves an R@1 of 67.10 with audio-only embeddings, compared with 57.30 under joint encoding, with similar declines in R@5, R@10, and MRR. With the query and filtered candidate set held fixed, this gap suggests that incorporating visual information can weaken alignment with audio semantics. Thus, **visual bias may reflect not only limited speech understanding but also interference from joint encoding** that makes otherwise retrievable audio-semantic information less accessible.

4.2.4 Qualitative Results

Figure 4 presents qualitative retrieval results from WAVE and Omni-Embed-Nemotron on our proposed AVEmbedBench. We randomly select a query and visualize its positive, audio-negative, and visual-negative videos. To investigate the retrieval ranking, we decompose the query into visual and audio clauses and compute their similarities to the corresponding video modalities, alongside the joint query-video similarity. In Figure 4(a), WAVE ranks the audio-negative video first and the positive second. Despite its unrelated speech, the negative has higher visual similarity (0.4174 vs. 0.3996) and only slightly lower audio similarity (0.3414 vs. 0.3464), yielding a higher joint score (0.4365 vs. 0.4256). Conversely, the visual-negative video ranks fifth despite having the highest audio similarity among the displayed candidates (0.3513). A similar pattern emerges for Omni-Embed-Nemotron in Figure 4(b): the audio-negative video ranks first, whereas the positive ranks fifth despite better audio alignment (0.4257 vs. 0.4139). The negative instead exhibits stronger visual alignment (0.4233 vs. 0.3860), accompanied by a higher joint score (0.4522 vs. 0.4190).

These two examples further illustrates a **visual-dominance bias** in omni-embedding models: **strong visual alignment can outweigh an audio-semantic mismatch**, highlighting the difficulty of distinguishing visually similar negatives through audio semantics in omni-modal retrieval.

5 Limitations and Future Work

Despite its value in assessing joint audio–visual retrieval, AVEmbedBench has two main limitations.

First, it focuses exclusively on text-to-video retrieval and thus does not fully capture the broader capabilities of multimodal embedding models. Future work will expand the benchmark to include

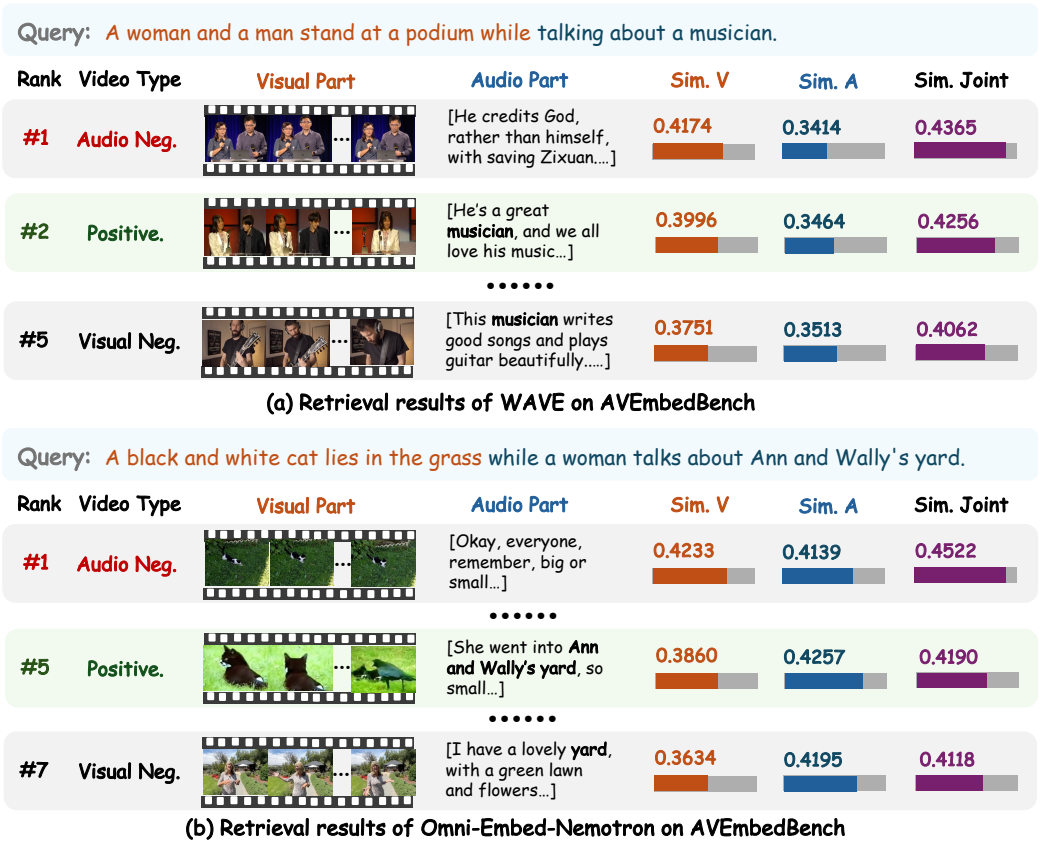


Figure 4: **Qualitative examples of visual bias in audio–visual joint retrieval on AVEmbed-Bench.** Results from (a) WAVE and (b) Omni-Embed-Nemotron show that visually aligned but audio-mismatched negatives rank above the positive videos. "Sim. V", "Sim. A", and "Sim. Joint" denote visual, audio, and joint query–video similarity scores, respectively. "Rank" indicates the ranking by joint similarity.

more diverse modality combinations and tasks, such as image/video question answering and video classification, enabling a more comprehensive assessment of joint audio–visual understanding.

Second, although AVEmbedBench reveals limitations in joint audio–visual retrieval, its potential to guide model training remains unexplored. Future work will leverage its contrastive samples for fine-tuning, encouraging models to align queries with videos that match both modalities while distinguishing them from negatives that match only one. We will explore parameter-efficient methods, such as LoRA and adapters, to strengthen speech-semantic representations, mitigate visual bias, and enhance joint audio–visual understanding in omni-modal embedding models.

6 Conclusion

In this paper, we introduce AVEmbedBench, to our knowledge the first benchmark specifically designed for joint audio–visual retrieval in omni-embedding models. By combining queries that impose both visual and speech-semantic constraints with hard negatives matching only one modality, AVEmbedBench challenges models to distinguish joint relevance from unimodal similarity. We develop an automated construction pipeline, yielding 275,409 retrieval instances spanning 637,481 unique video clips. Our evaluation reveals a consistent visual bias: models tend to favor visually similar candidates despite mismatched speech content. These findings highlight the gap between supporting multimodal inputs and effectively integrating them, motivating more balanced audio–visual integration for reliable joint retrieval.

AI USE STATEMENT

We disclose our use of generative AI tools along the categories defined by the ICLR 2027 AI policy. **Required disclosure.** Generating synthetic datasets: not used; AVEmbedBench was constructed by selecting and extracting data from InternVideo2_Vid_Text without AI assistance. Helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in writing proofs: not used. Proposing or refining hypotheses: not used. Designing or providing feedback on research methodology or experiments: not used. Implementing methods: not used. Assisting with translation: not used. Cleaning and reformatting datasets: not used. Supporting qualitative and thematic data analysis: not used. Interpreting results: not used. **Recommended disclosure.** Drafting parts of the research paper and editing it to improve readability: used; generative AI tools assisted in drafting and polishing the manuscript text. Formatting references: not used. We have reviewed all AI-assisted work. The authors take responsibility for the final content of this work, including its text, claims, and artifacts.

ETHICS STATEMENT

This work evaluates visual–audio joint retrieval in omni-embedding models using AVEmbedBench, constructed from existing publicly available video datasets, and involves no human-subject experiments or newly collected video recordings. All video clips are selected from these sources, and retrieval performance is assessed using automated metrics rather than human judges. Because the benchmark uses existing videos and evaluates pretrained models without modifying their parameters, the results may reflect the behaviors, limitations, and potential biases of the source data and evaluated models; our experiments do not assess broader ethical biases, and we make no claim that AVEmbedBench reduces them. The accompanying analysis focuses on retrieval performance and modality bias and does not characterize broader fairness, privacy, or safety, so it should not be read as a guarantee regarding these dimensions. Reported results are limited to the models, data, and retrieval settings studied here, and the qualitative examples illustrate individual retrieval outcomes; they should not be generalized to other models, domains, or deployment settings, nor interpreted as evidence of broader safety or reliability improvements.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we document the data source and construction pipeline of AVEmbedBench, including video filtering, joint query generation, and modality-specific hard negative mining. We describe the evaluated models, retrieval protocols, and evaluation metrics, and provide dataset statistics in Table 1. For evaluation without hard negatives, we remove only the two negatives associated with each query while keeping all other candidates unchanged, ensuring a consistent and reproducible comparison. We commit to publicly releasing AVEmbedBench upon publication.

REFERENCES

Ashutosh Chaubey, Jiacheng Pang, and Mohammad Soleymani. MoD-DPO: Towards mitigating cross-modal hallucinations in omni LLMs using modality decoupled preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Haonan Chen, Sicheng Gao, Radu Timofte, Tetsuya Sakai, and Zhicheng Dou. e5-omni: Explicit cross-modal alignment for omni-modal embeddings. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19430–19443, 2026a.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Chao Sun, Rongzhou Zhang, Guanyu Zhou, and Lijie Wen. OmniDPO: A preference optimization framework to address omni-modal hallucination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 20172–20180, 2026b.

Sangyun Chung, Se Yeon Kim, Youngchae Chee, and Yong Man Ro. MAD: Modality-adaptive decoding for mitigating cross-modal hallucinations in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 736–740, 2020.

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

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

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representations*, 2025.

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

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

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. In *Advances in Neural Information Processing Systems*, 2025.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. VLM2Vec-V2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025.

Microsoft. Harrier-OSS-v1-27B. Hugging Face model card, 2026.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, and James Glass. AVLnet: Learning audio-visual language representations from instructional videos. In *Proceedings of Interspeech*, pp. 1584–1588, 2021. doi: 10.21437/Interspeech.2021-1312.

Madhuri Shanbhogue, Zhe Li, Shanfeng Zhang, et al. Gemini embedding 2: A native multimodal embedding model from gemini. *arXiv preprint arXiv:2605.27295*, 2026a.

Madhuri Shanbhogue, Zhe Li, Shanfeng Zhang, et al. Gemini Embedding 2: A native multimodal embedding model from Gemini. *arXiv preprint arXiv:2605.27295*, 2026b.

Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S. Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once - multi-modal fusion transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20020–20029, 2022.

Changli Tang, Qinfan Xiao, Ke Mei, Tianyi Wang, Fengyun Rao, and Chao Zhang. Wave: Learning unified & versatile audio-visual embeddings with multimodal llm. In *Proceedings of the International Conference on Learning Representations*, 2026.

Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. ONE-PEACE: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: A systematic review of methods and future directions. *Proceedings of the IEEE*, 112 (11):1716–1754, 2024a. doi: 10.1109/JPROC.2024.3525147.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *Proceedings of the European Conference on Computer Vision*, pp. 396–416, 2024b.

Chenghao Xiao, Hou Pong Chan, Hao Zhang, Weiwen Xu, Mahani Aljunied, and Yu Rong. Scaling language-centric omnimodal representation learning. In *Advances in Neural Information Processing Systems*, 2025a.

Chenghao Xiao, Hou Pong Chan, Hao Zhang, Weiwen Xu, Mahani Aljunied, and Yu Rong. Scaling language-centric omnimodal representation learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 158370–158401, 2025b.

Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman Solomatin, Noura Al Moubayed, Kenneth Enevoldsen, and Niklas Muennighoff. Mieb: Massive image embedding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22187–22198, 2025c.

Mengyao Xu, Wenfei Zhou, Yauhen Babakhin, Gabriel Moreira, Ronay Ak, Radek Osmulski, Bo Liu, Even Oldridge, and Benedikt Schifferer. Omni-embed-nemotron: A unified multimodal retrieval model for text, image, audio, and video. *arXiv preprint arXiv:2510.03458*, 2025a.

Mengyao Xu, Wenfei Zhou, Yauhen Babakhin, Gabriel Moreira, Ronay Ak, Radek Osmulski, Bo Liu, Even Oldridge, and Benedikt Schifferer. Omni-Embed-Nemotron: A unified multimodal retrieval model for text, image, audio, and video. *arXiv preprint arXiv:2510.03458*, 2025b.

Chenshuang Zhang, Kyeong Seon Kim, Chengxin Liu, and Tae-Hyun Oh. Svhalluc: Benchmarking speech-vision hallucination in audio-visual large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25304–25314, 2026.