

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 OMNICVR: A BENCHMARK FOR OMNI-COMPOSED VIDEO RETRIEVAL WITH VISION, AUDIO, AND TEXT

Anonymous authors

Paper under double-blind review

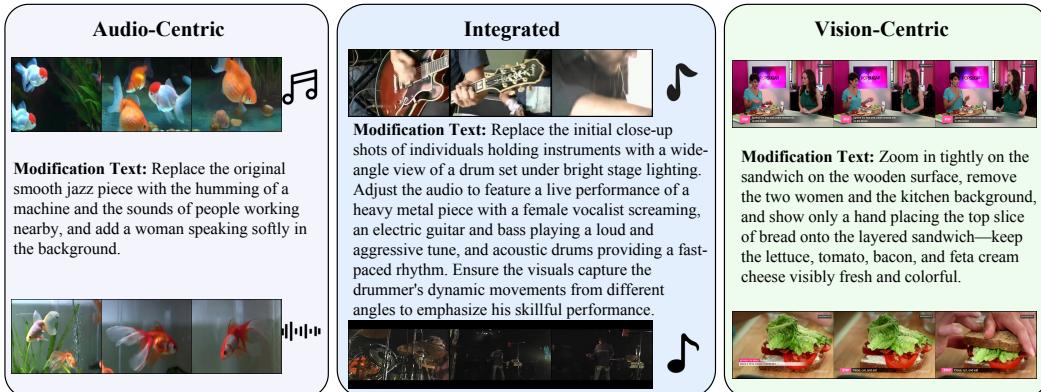


Figure 1: Overview of the OmniCVR Benchmark.

## ABSTRACT

Composed video retrieval presents a complex challenge: retrieving a target video based on a source video and a textual modification instruction. This task demands fine-grained reasoning over multimodal transformations. However, existing benchmarks predominantly focus on vision–text alignment, largely overlooking the rich semantic signals embedded in audio—such as speech, music, and environmental sounds—which are often decisive for comprehensive video understanding. To bridge this gap, we introduce **OmniCVR**, a large-scale benchmark for omni-composed video retrieval that establishes vision, audio, and text as first-class modalities. OmniCVR is constructed via a scalable, automated pipeline integrating content-aware segmentation, omni-modal annotation, and a rigorous dual-validation protocol involving both large language models and human experts. The benchmark comprises vision-centric, audio-centric, and integrated queries, with the latter forming the majority to accurately reflect real-world multimodal complexity. Furthermore, we propose **AudioVLM2Vec**, an audio-aware extension of VLM2Vec. By incorporating explicit audio semantics, AudioVLM2Vec achieves state-of-the-art performance, highlighting fundamental limitations in the audio reasoning capabilities of current multimodal retrieval systems.

## 1 INTRODUCTION

Video has established itself as the dominant medium for global communication, education, and entertainment, precipitating an exponential growth in digital video content. This deluge has necessitated retrieval systems capable of searching vast repositories with both accuracy and efficiency. While early content-based retrieval depended on low-level visual features, the advent of large-scale vision–language models (Zhan et al., 2024; Kelly et al., 2024; Peng et al., 2024) has revolutionized the field, enabling robust text-to-video retrieval (Radford et al., 2021). Foundational benchmarks such as MSR-VTT (Xu et al., 2016), VATEX (Wang et al., 2019), and YouCook2 (Zhou et al., 2018)

054 have been instrumental in this evolution, pairing extensive video collections with natural language  
 055 captions to advance video–language alignment.

056 Recently, the paradigm has shifted toward *composed video retrieval* (CoVR), which requires models  
 057 to retrieve a target video given a source video and a specific textual modification instruction (Thawakar  
 058 et al., 2024; Hummel et al., 2024; Thawakar et al., 2025). This formulation demands not only visual  
 059 grounding but also fine-grained compositional reasoning—for instance, “retrieve the same cooking  
 060 scene but with a different ingredient.” Such benchmarks have successfully pushed the boundary from  
 061 simple retrieval toward complex reasoning tasks.

062 Despite these advancements, a critical limitation remains: existing benchmarks overwhelmingly treat  
 063 video as a purely visual–textual medium, neglecting the audio stream. Audio often carries semantic  
 064 weight equal to or greater than vision; speech conveys intent, background music establishes mood,  
 065 and environmental sounds define context. A scene depicting “a crowd cheering at a sports arena” is  
 066 incompletely represented by visuals alone. By ignoring audio, current benchmarks fail to evaluate  
 067 models in scenarios where auditory information is decisive. Furthermore, no existing framework  
 068 systematically addresses retrieval tasks requiring simultaneous modifications across both vision and  
 069 audio.

070 To address this deficiency, we present **OmniCVR: the first benchmark for omni-composed video**  
 071 **retrieval**, treating vision, audio, and text as unified, first-class modalities. OmniCVR introduces  
 072 large-scale, compositional retrieval tasks spanning three distinct categories, **as illustrated in Figure 1**:  
 073 vision-centric (modifying actions, objects, or scenes), audio-centric (altering music, sound effects, or  
 074 speech while preserving visual similarity), and integrated (simultaneously modifying both modalities).  
 075 Unlike prior works, integrated queries dominate OmniCVR, reflecting the intricate multimodal nature  
 076 of real-world video. The benchmark is constructed via a scalable automated pipeline that combines  
 077 segmentation, omni-modal annotation, and a dual-validation mechanism (utilizing Gemini 2.5 Pro  
 078 and human experts in an AND-gated protocol) to ensure both breadth and high-quality data.

079 In summary, our primary contributions are:

- 081 We introduce **OmniCVR**, the inaugural large-scale benchmark for omni-modal composed  
 082 video retrieval, comprising 50K triplets derived from 160K clips and a rigorously validated  
 083 5K-instance gold-standard test set.
- 084 We propose a scalable data generation pipeline integrating content-aware video segmen-  
 085 tation, omni-modal annotation, and dual validation, yielding high-quality, compositional  
 086 instructions.
- 087 We evaluate seven baselines and propose **AudioVLM2Vec**, which achieves state-of-the-art  
 088 results on OmniCVR, revealing significant gaps in existing methods regarding audio-centric  
 089 and compositional reasoning.

## 091 2 RELATED WORK

### 093 2.1 VIDEO-TEXT RETRIEVAL BENCHMARKS

095 The cornerstone of modern video retrieval research is the availability of large-scale video-text  
 096 datasets. MSR-VTT Xu et al. (2016) pioneered this space with 10,000 web video clips, followed by  
 097 VATEX Wang et al. (2019), which expanded the scale to over 41,000 clips with bilingual captions.  
 098 Domain-specific datasets such as YouCook2 Zhou et al. (2018) (instructional cooking) and Charades  
 099 Sigurdsson et al. (2016) (indoor activities) further diversified the field. **Similarly, recent works like**  
 100 **MultiVENT 2.0 Kriz et al. (2025) have scaled retrieval to massive multilingual and event-centric do-**  
 101 **mains.** While these benchmarks have been instrumental in advancing video–language understanding,  
 102 they predominantly focus on visual content, largely neglecting the auditory modality. **Consequently,**  
 103 **they fail to incorporate Composed Video Retrieval scenarios involving audio modifications, such as**  
 104 **keeping the visual scene but changing the background music.**

### 105 2.2 COMPOSED VIDEO RETRIEVAL BENCHMARKS

106 To transcend simple text-based retrieval, the task of Composed Video Retrieval was introduced (Ven-  
 107 tura et al., 2024; Gupta et al., 2025; Yue et al., 2025). CoVR tasks a model with retrieving a target

108 video given a source video (or image) and a textual instruction detailing the desired modification.  
 109 **WebVid-CoVR** Thawakar et al. (2024) established a large-scale, synthetic dataset for this purpose.  
 110 Subsequently, **Dense-WebVid-CoVR** Thawakar et al. (2025) utilized GPT-4o to generate more  
 111 elaborate modification texts. In the egocentric domain, **EgoCVR** Hummel et al. (2024) focused on  
 112 subtle, temporal, and action-oriented modifications. As detailed in Figure 2 and Table 1, a recurrent  
 113 limitation among these CoVR benchmarks is their exclusive focus on visual modifications, leaving the  
 114 auditory dimension unexplored. OmniCVR is the first to introduce compositional queries involving  
 115 audio changes, such as “change the background music to an upbeat pop track.”

116

117 Table 1: Comparison of OmniCVR with existing video retrieval benchmarks. OmniCVR is the first  
 118 to explicitly incorporate a detailed, searchable audio modality and support composed audio-visual  
 119 queries.

120

Benchmark	Data Source	Annotation Method	Scale	Task Focus	Audio Modality	CVR
MSR-VTT Xu et al. (2016)	Web Videos	Crowdsourced	10K clips	Text-to-Video	✓	✗
VATEX Wang et al. (2019)	Web Videos	Crowdsourced	41K clips	Text-to-Video	✓	✗
WebVid-CoVR Thawakar et al. (2024)	WebVid10M	Synthetic (LLM)	470K triplets	Visual Comp.	✗	✓
EgoCVR Hummel et al. (2024)	Ego4D	Manual	2.3K queries	Temporal Comp.	✗	✓
Dense-WebVid-CoVR Thawakar et al. (2025)	WebVid	Synthetic (GPT-4o)	1.6M samples	Fine-grained Vis.	✗	✓
<b>OmniCVR (Ours)</b>	Diverse	Generative (Qwen-Omni)	160K+ clips	<b>Omni-Composed</b>	✓	✓

125

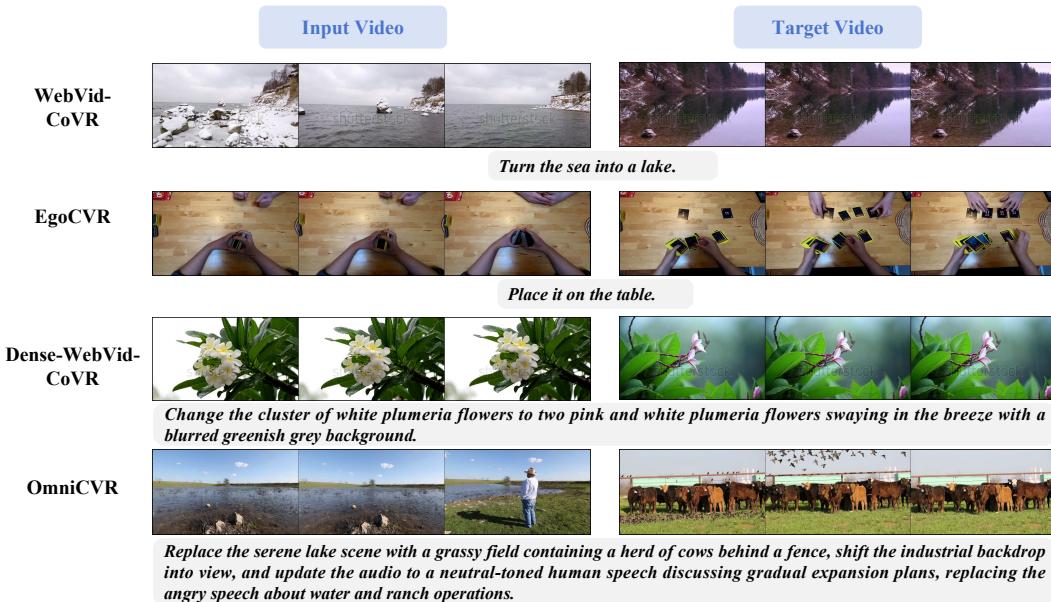


Figure 2: Comparison of OmniCVR with existing video retrieval benchmarks.

### 2.3 AUDIO–VISUAL LEARNING

Beyond video–text retrieval, there is growing interest in joint vision-audio learning. Tasks such as *audio–visual source separation* Afouras et al. (2018) and *audio–visual event localization* Tian et al. (2018) demonstrate the efficacy of multimodal fusion. Benchmarks like **AV-SUPERB** Tseng et al. (2024) evaluate multi-task representations across speech and sound, while recent long-video benchmarks like **MLVU** Zhou et al. (2025) emphasize the challenges of integrating extended temporal and multimodal information. However, these efforts are restricted primarily to classification or grounding; none provide *compositional retrieval tasks* that explicitly require models to adhere to natural language instructions spanning both audio and vision.

162 

### 3 THE OMNICVR BENCHMARK

164 

#### 3.1 OVERVIEW OF OMNICVR

166 We introduce the Omni Composed Video Retrieval Benchmark (OmniCVR), a large-scale omni-modal  
 167 framework designed to evaluate the compositional retrieval capabilities of multimodal foundation  
 168 models. Unlike prior CoVR benchmarks focused solely on vision, OmniCVR systematically integrates  
 169 **video, audio, and text**. Each instance is represented as a triplet: *(source video, modification text,*  
 170 *target video)*, where the text specifies the transformation required to map the source to the target.

171 We curated over **50,000 original long-form videos**, segmenting them into **160,000 coherent short**  
 172 **clips**. These clips are annotated and paired to produce **50,000 compositional triplets**. From this  
 173 corpus, a **5,000-triplet gold-standard test set** was selected and manually validated. OmniCVR tasks  
 174 fall into three categories:

- 176 • **Vision-Centric:** Queries focusing on modifying actions, objects, or scenes.
- 177 • **Audio-Centric:** Queries altering music, sound effects, or speech while maintaining high  
 178 visual similarity.
- 179 • **Integrated:** Queries requiring simultaneous modifications across both visual and auditory  
 180 modalities.

182 This design ensures OmniCVR evaluates not only perceptual grounding but also the capacity to  
 183 follow complex multimodal modification instructions.

185 Table 2: Core statistics of OmniCVR. Integrated queries dominate the benchmark, reflecting the  
 186 focus on realistic cross-modal modifications.

188 Statistics	189 Number
189 Training Triplets	45k+
190 Unique Video Clips	160k+
191 Test Set (gold-standard)	5,000
192 Query Types (Vision:Audio:Integrated)	22.82% : 20.00% : 57.18%
193 Avg. Query Length	52.6 words
194 Vocabulary Size	25k+
195 Avg. Video Length	11.8 sec

197 Table 2 summarizes the benchmark statistics. Unlike existing datasets, OmniCVR emphasizes  
 198 **integrated queries**, which constitute the majority of tasks. This distribution mirrors real-world  
 199 complexity, where modifications rarely occur in isolation. Vision-centric queries form the second  
 200 largest cohort, while audio-centric queries are rarer due to strict pairing constraints (high visual  
 201 similarity, low audio similarity). It is worth noting that the average query length in OmniCVR (52.6  
 202 words) is higher than in visual-only CVR benchmarks. This is a deliberate design choice: Integrated  
 203 queries (57.18% of the dataset) require specifying simultaneous transformations in both visual and  
 204 auditory domains to avoid ambiguity. Unlike prior works that overlook audio, OmniCVR necessitates  
 205 denser descriptions to capture the full spectrum of multimodal evolution, reflecting the complexity of  
 206 real-world video editing and retrieval scenarios.

207 To further demonstrate the semantic breadth of the benchmark, Figure 3 visualizes the hierarchical dis-  
 208 tribution of content across modalities. Regarding video content (Figure 3a), OmniCVR encompasses  
 209 four primary domains: *Entertainment & Events, Instructional & Procedural, Daily Life, Nature &*  
*210 Travel*, and *Music & Performance*. These are structured into granular subcategories—ranging from  
 211 *Culinary Arts* to *Wilderness & Wildlife*—to ensure robust coverage of diverse visual scenes. Similarly,  
 212 the auditory landscape is rigorously balanced, as shown in Figure 3b. We organize audio instances  
 213 into a two-level taxonomy comprising three high-level classes (*Speech, Music, and Sound*) branched  
 214 into 15 distinct subcategories (e.g., *Scripted Dialogue, Instrumental Music, and Nature & Animal*  
 215 *Sounds*). This fine-grained categorization highlights the acoustic richness and real-world complexity  
 embedded in our benchmark.

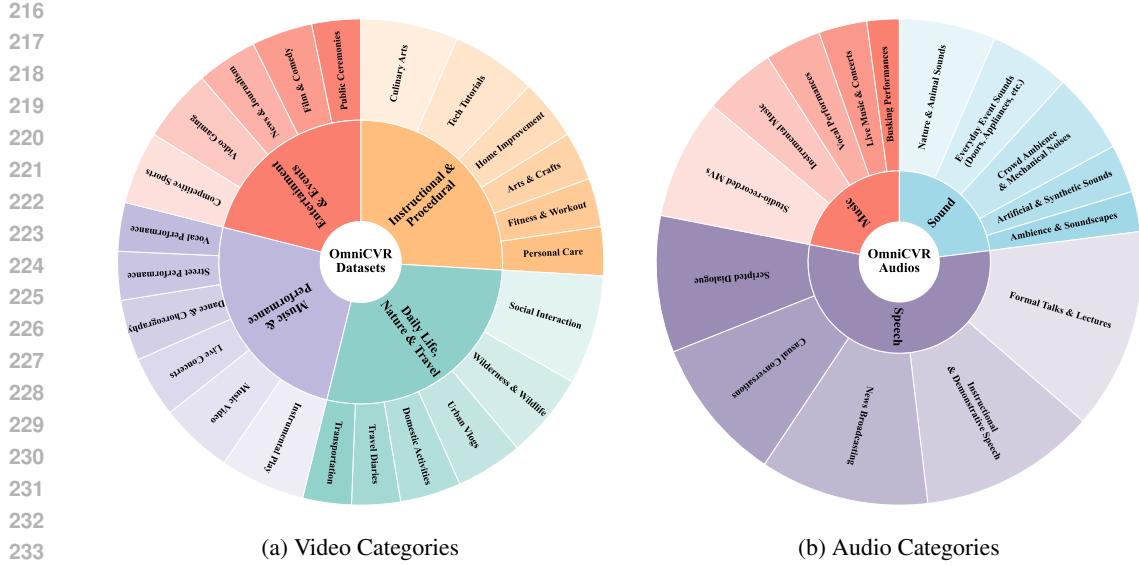


Figure 3: **The data diversity of OmniCVR.** (a) Illustrates the hierarchical taxonomy of video content, spanning four major domains and their subcategories. (b) Displays the distribution of audio types, categorized into Speech, Music, and Sound with fine-grained subdivisions.

### 3.2 DATA GENERATION PIPELINE

OmniCVR is constructed via a three-stage automated pipeline, depicted in Figure 4, ensuring both scale and high fidelity.

**Stage 1: Video Curation and Segmentation.** We begin by collecting long-form videos (typically ranging from several minutes to multiple hours) from a diverse set of public datasets, including HowTo100M (Miech et al., 2019), MSR-VTT, VATEX, and others. These sources cover diverse domains such as instructional content, daily activities, and entertainment.

From these long-form videos, we extract semantically coherent short clips (5–15 seconds, average 11.8 seconds) that serve as the final retrieval units. We employ PySceneDetect to define segments based on inter-frame HSV differences ( $\tau = 36$ ), effectively capturing scene changes while ignoring minor camera motion. Post-segmentation, we filter clips using two metrics: **Action Intensity** (optical flow magnitude) and **Scene Richness** (visual feature variance). Only clips exceeding a combined threshold are retained, ensuring semantic density.

**Stage 2: Generative Omni-Modal Annotation.** Clips are annotated using Owen2.5-Omni, which jointly encodes video and audio. To enhance auditory detail, we integrate automatic **audio transcription**. The annotation prompt requests structured descriptions of scenes, actions, objects, and audio events. Specifically for audio, to capture the full acoustic spectrum, we enforce a strict schema covering **para-linguistic features, lexical content, environmental hierarchy, and temporal dynamics** (see Appendix G for full details). These dimensions are validated in Stage 3 to ensure retrieval is grounded in fine-grained audio semantics. Quality is enforced through a two-stage verification process: automatic consistency checks with Gemini 2.5 Pro and manual expert review.

**Stage 3: Triplet Mining for Compositional Retrieval.** We generate triplets via three strategies:

- **Vision-Centric:** Constructed either from (a) different segments of the same long video, or (b) distinct clips from the same video source, ensuring coverage of both coarse-grained and fine-grained visual differences. Audio is preserved to isolate visual reasoning.
- **Audio-Centric:** Candidate pairs are first filtered by requiring high visual similarity (video CLIP cosine similarity  $> 0.9$ ). Among these, pairs with low auditory similarity (audio

embedding cosine similarity  $< 0.3$ , measured by the CLAP model) are selected. This guarantees that the visual scene remains constant while audio varies significantly.

- **Integrated:** Pairs are chosen to differ in both modalities, with low similarity in CLIP embeddings (vision) and low similarity in CLAP embeddings (audio).

Modification texts are generated by prompting an LLM with the structured annotations of the source and target, explicitly encoding the relevant differences.

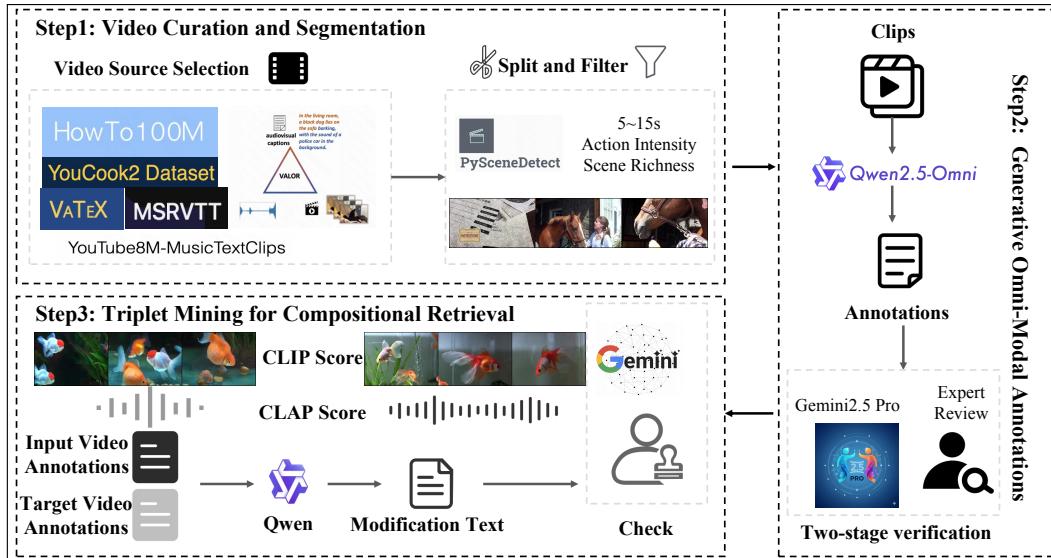


Figure 4: OmniCVR Benchmark Construction Pipeline.

### 3.3 EVALUATION AND TEST SET CURATION

We construct a gold-standard test set of **5,000** instances using a *concurrent dual-gate protocol*. For every candidate triplet, both **Gemini 2.5 Pro** and a **human expert** independently validate the paired videos and modification text. A triplet is admitted only if *both* approve. This AND-gated review guarantees semantic fidelity and consistency. The resulting test set preserves the natural query distribution (Integrated  $>$  Vision  $>$  Audio), creating a realistic evaluation regime.

Table 3: Distribution of audio categories (speech, music, sound) in the source and target videos.

Video Type	Speech (%)	Music (%)	Sound (%)
Source video	56	23	21
Target video	55	23	22

Table 3 illustrates the modality distribution of the audio streams in the source and target videos. We decompose each video’s soundtrack into three categories: *speech*, *music*, and *sound*.

## 4 EXPERIMENTAL SETUP

### 4.1 TASKS AND EVALUATION

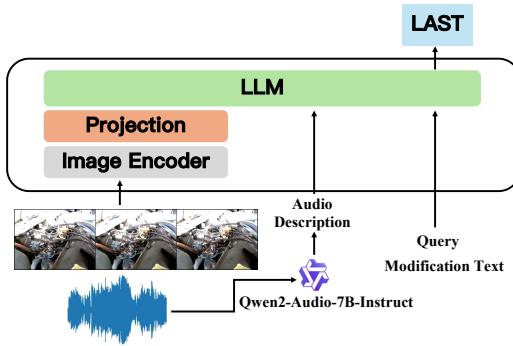
OmniCVR is designed to rigorously assess compositional retrieval across multiple modalities, focusing on **Composed Video Retrieval (CVR)**. In this task, a model is given a source video and a natural language instruction and must retrieve the corresponding target video from a candidate pool. Unlike traditional video retrieval, which emphasizes semantic similarity, CVR tests a model’s ability to reason over transformations like object changes or action alterations. Queries are categorized

324 into three groups: vision-only modifications (focused on visual appearance and motion), audio-only  
 325 modifications (based on acoustic cues like speech or background music), and joint vision–audio  
 326 modifications (requiring integrated multimodal reasoning). Performance is measured using standard  
 327 retrieval metrics—Recall at K (R@1, R@5, R@10)—which reflect the accuracy of retrieving the  
 328 correct target from the top candidates. This provides a challenging evaluation that mirrors real-world  
 329 multimodal retrieval tasks.

## 331 4.2 BASELINE MODELS

333 **Large Multimodal Embedding Models.** (i) **OmniEmbed-v0.1-multivent** (Ma et al., 2025): A  
 334 unified model from Tevatron 2.0 that encodes text, image, audio, and video into a shared space,  
 335 achieving state-of-the-art performance in cross-modal video retrieval, particularly on the MAGMaR  
 336 Shared Task (Zhan et al., 2025), after fine-tuning on MultiVENT data (Kriz et al., 2025) with joint  
 337 vision–audio–text supervision. (ii) **VLM2Vec** (Jiang et al., 2025): A framework that converts large  
 338 vision–language models (e.g., Qwen2–VL (Wang et al., 2024), LLaVA (Liu et al., 2023), Phi-3.5–V)  
 339 into universal embedding models through contrastive learning on the MMEB benchmark. VLM2Vec  
 340 outperforms baselines like CLIP (Radford et al., 2021) and BLIP (Li et al., 2022), showing 10–  
 341 20% improvement in retrieval tasks. (iii) **AudioVLM2Vec (Ours)**: Our extension of VLM2Vec  
 342 that integrates audio semantics by using **Qwen2–Audio** (Chu et al., 2024) to generate fine-grained  
 343 captions of video audio tracks. The captions are combined with user queries and fed into VLM2Vec,  
 344 enhancing its representation power while focusing on audio signals.

345 **Lightweight and Task-Specific Models.** (iv) **CLIP** (Radford et al., 2021): A foundational vision–  
 346 language model trained on large image–text pairs, adapted for video retrieval via frame-level temporal  
 347 averaging. (v) **CoVR** (Thawakar et al., 2024): A video retrieval model optimized for vision–centric  
 348 modifications, evaluated for its generalization beyond visual changes. (vi) **BLIP** (Li et al., 2022): A  
 349 unified vision–language model adapted for retrieval tasks using its vision–language matching head.  
 350 (vii) **ImageBind** (Girdhar et al., 2023): A model learning a joint embedding across six modalities,  
 351 including audio and video, ideal for audio–centric tasks.



364 Figure 5: The framework of AudioVLM2Vec.

365 Figure 5 illustrates the design of our proposed **AudioVLM2Vec** model. The framework extends  
 366 VLM2Vec by explicitly injecting audio semantics into the retrieval pipeline. Given a video, we first  
 367 encode its visual content using a pretrained image encoder followed by a lightweight projection layer.  
 368 In parallel, the audio track is processed by **Qwen2–Audio–7B–Instruct**, which generates a fine-grained  
 369 natural language description of the acoustic scene. This audio-derived text is concatenated with the  
 370 user’s modification query and then fed into the LLM backbone, ensuring that both vision and audio  
 371 cues are aligned in a shared semantic space. **By feeding the transcribed audio semantics alongside  
 372 visual tokens into the LLM, we leverage the model’s multi-head self-attention mechanism to jointly  
 373 process both modalities. This allows the model to learn synergistic and causal relationships (e.g.,  
 374 aligning the text ‘lips moving’ with corresponding visual tokens) within a shared high-dimensional  
 375 semantic space.** The final multimodal embedding is obtained from the LLM and optimized for  
 376 retrieval via contrastive learning. By translating audio signals into text and integrating them at the  
 377 embedding stage, AudioVLM2Vec effectively grounds compositional queries in both modalities,  
 378 yielding substantial gains on audio–centric and cross-modal retrieval tasks.

378 4.3 EVALUATION STRATEGY  
379380 For each query, we compute similarity scores between the query embedding and candidate video  
381 embeddings, ranking candidates accordingly. To mitigate potential variance introduced by candidate  
382 set composition, we shuffle candidate pools five times and report averaged metrics. For audio-  
383 centric tasks, we additionally control for modality imbalance by ensuring candidate pools always  
384 include visually similar but acoustically distinct distractors, and vice versa. This evaluation protocol  
385 ensures fairness across different model families while highlighting the challenges of multimodal  
386 compositional retrieval.  
387388 5 RESULTS AND DISCUSSION  
389390 5.1 MAIN RESULTS  
391392 Table 4 presents Recall@K on OmniCVR across overall queries, while Table 5 isolates the case of  
393 *audio-centric queries*. We summarize the main findings below. **The symmetric vision-centric retrieval**  
394 **results on OmniCVR are reported in Table 10 in Appendix C.**395 1. **Audio-centric queries expose a critical failure mode in existing baselines.** While our  
396 AudioVLM2Vec adapts robustly to audio-dependent queries (77.2 R@1), strong baselines  
397 suffer catastrophic performance degradation. For instance, VLM2Vec drops from an overall  
398 R@1 of 38.44 to just 12.4 in the audio-centric setting. This disparity underscores the unique  
399 difficulty of audio compositionality compared to visual tasks, a challenge that prior models  
400 fail to address.  
401 2. **Large-scale multimodal models establish a superior performance tier.** Across all query  
402 types, large embedding models consistently and substantially outperform lightweight, task-  
403 specific retrievers, validating the efficacy of large-scale pre-training for compositional  
404 reasoning.  
405 3. **AudioVLM2Vec achieves universal state-of-the-art performance.** Our model ranks first  
406 across all categories and  $K$  values, achieving **66.98** overall R@1 and **77.2** on audio-centric  
407 queries. This consistency confirms that our architecture generalizes effectively across both  
408 unimodal and integrated multimodal retrieval tasks.  
409 4. **Explicit audio semantics are the decisive factor for performance gains.** The impact  
410 of injecting audio descriptions is profound: AudioVLM2Vec surpasses the VLM2Vec  
411 baseline by an impressive **+64.8** absolute points (77.2 vs. 12.4) in the audio-centric setting  
412 and **+28.5** points overall. These results demonstrate that audio-aware embeddings are not  
413 merely beneficial but indispensable for handling compositional queries involving non-visual  
414 transformations.  
415416  
417 Table 4: Overall performance comparison of baseline models on OmniCVR. We report Recall at K  
418 (R@1, R@3, R@5, R@10). Best results within each group are highlighted in **bold**, and second-best  
419 are underlined.  
420

421 Models	422 Backbone	423 R@1	424 R@3	425 R@5	426 R@10
<b>Lightweight and Task-Specific Models</b>					
427 CLIP	428 CLIP	429 27.54	430 50.46	431 56.70	432 62.62
433 CoVR	434 BLIP2 (Li et al., 2023)	435 11.46	436 22.88	437 28.08	438 35.18
439 BLIP	440 BLIP	441 6.3	442 11.84	443 14.12	444 17.00
446 IMAGEBIND	447 CLIP	448 17.28	449 29.55	450 43.34	451 45.33
<b>Large Multimodal Embedding Models</b>					
453 OmniEmbed-v0.1-multivent	454 Qwen2.5-Omni	455 31.90	456 51.50	457 57.04	458 64.00
459 VLM2Vec	460 Qwen2-VL	461 <u>38.44</u>	462 <u>55.48</u>	463 <u>60.44</u>	464 <u>66.60</u>
466 AudioVLM2Vec (Ours)	467 Qwen2-Audio + Qwen2-VL	468 <b>66.98</b>	469 <b>77.84</b>	470 <b>80.86</b>	471 <b>84.40</b>

432 Table 5: Audio-centric retrieval performance of Large Multimodal Embedding Models on OmniCVR.  
 433 We report Recall at K (R@1, R@3, R@5, R@10). Best results are highlighted in **bold**.

435 <b>Models</b>	436 <b>Backbone</b>	437 R@1	438 R@3	439 R@5	440 R@10
437 OmniEmbed-v0.1-multivent	438 Qwen2.5-Omni	439 13.6	440 28.5	441 35.8	442 47.0
438 VLM2Vec	439 Qwen2-VL	440 12.4	441 23.3	442 30.4	443 42.3
439 AudioVLM2Vec (Ours)	440 Qwen2-Audio + Qwen2-VL	441 <b>77.2</b>	442 <b>87.3</b>	443 <b>90.7</b>	444 <b>94.2</b>

## 441 5.2 DETAILED ANALYSIS AND ABLATION STUDIES

443 **Breakdown of Audio-Centric Performance.** To better understand the model’s capabilities across  
 444 different acoustic domains, we decompose the audio-centric performance by target audio category  
 445 (Human Speech, Music, and Sound). As shown in Table 6, AudioVLM2Vec achieves dominant  
 446 performance in Human Speech (+85.23% gain) and Music (+70.36% gain). This indicates that  
 447 converting audio to text effectively captures both lexical content and para-linguistic features like  
 448 genre and mood. The gain is slightly lower but still substantial for Sounds (+49.56%), reflecting the  
 449 inherent challenge in describing unstructured acoustic events compared to structured speech or music.

450 Table 6: Fine-grained breakdown of audio-centric retrieval (R@1) on OmniCVR by target audio  
 451 category. Best results are highlighted in **bold**.

454 <b>Target Audio Category</b>	455 <b>VLM2Vec</b>	456 <b>AudioVLM2Vec (Ours)</b>	457 <b>Absolute Gain</b>
455 Human Speech	456 11.36	457 <b>96.59</b>	458 +85.23
456 Music	457 16.07	458 <b>86.43</b>	459 +70.36
457 Sound	458 10.75	459 <b>60.31</b>	460 +49.56

461 **Impact of Source Video.** To determine if detailed modification texts render the source video  
 462 redundant (effectively reducing the task to text-to-video retrieval), we conducted a “Blind Retrieval”  
 463 ablation. Here, we removed the visual frames of the source video, forcing the model to rely  
 464 solely on the modification instruction and source audio. As shown in Table 7, performance drops  
 465 catastrophically without the source video. For AudioVLM2Vec on audio-centric queries, R@1  
 466 plummets by 49.1% (from 77.20% to 28.10%). This confirms that the modification text functions as a  
 467 relative instruction rather than a standalone description. The source video provides essential context  
 468 (e.g., the visual scene “park”) to filter distractors, validating that OmniCVR rigorously evaluates  
 469 compositional reasoning.

470 Table 7: Ablation study on the importance of source video for VLM2Vec on OmniCVR audio-centric  
 471 retrieval.

472 <b>Metric</b>	473 <b>VLM2Vec (With Source Video)</b>	474 <b>Blind / Text-Only (No Source Video)</b>	475 <b>Performance Drop (<math>\Delta</math>)</b>
473 R@1	474 <b>77.20%</b>	475 28.10%	476 -49.10%
474 R@3	475 <b>87.30%</b>	476 33.20%	477 -54.10%
475 R@5	476 <b>90.70%</b>	477 42.50%	478 -48.20%
476 R@10	477 <b>94.20%</b>	478 57.80%	479 -36.40%

480 **Efficiency Analysis.** We benchmarked inference latency on an NVIDIA A800 GPU for 10-second  
 481 videos. While AudioVLM2Vec increases latency from 1.72s (VLM2Vec) to 4.77s (approx. 1.77x  
 482 increase in processing overhead due to audio transcription), this trade-off yields a 64.8% improvement  
 483 in audio-centric retrieval accuracy. Furthermore, with a Real-Time Factor (RTF) of 0.5, the system  
 484 remains faster than real-time playback, ensuring deployability.

485 **Native Audio Tokens vs. Audio-as-Text.** To rigorously isolate the contribution of our Audio-  
 486 as-Text fusion mechanism from backbone differences, we performed a controlled ablation on  
 487 **OmniEmbed-v0.1-multivent**. This model’s backbone (**Qwen2.5-Omni**) natively accepts raw audio  
 488 waveforms via a dedicated audio tower. We compared the original model (audio tower enabled)

486 against a modified version where the audio tower is disabled and replaced by our Qwen2-Audio-7B-  
 487 Instruct transcribed captions.  
 488

489 As shown in Table 8, simply replacing latent audio tokens with explicit textual descriptions—while  
 490 keeping the backbone, projector, and all training data identical—yields a dramatic improvement from  
 491 13.6 to **32.7** on R@1 (+19.1 absolute points, 2.4x relative gain).

492  
 493 Table 8: **Controlled ablation on OmniEmbed-v0.1-multivent backbone: native audio tokens vs. our**  
 494 **Audio-as-Text fusion mechanism on audio-centric retrieval. Best results are highlighted in bold.**

Model Setting	Audio Mechanism	R@1	R@3	R@5	R@10
OmniEmbed (Original)	Native Audio Tokens	13.6	28.5	35.8	47.0
OmniEmbed (Modified)	Audio-as-Text (Ours)	<b>32.7</b>	<b>48.0</b>	<b>58.9</b>	<b>69.1</b>

500 We further break down the results of this ablation by target audio category. As shown in Table 9, our  
 501 **Audio-as-Text strategy delivers consistent and substantial improvements across all categories**,  
 502 including a near-doubling of R@1 on Music (+13.93 points) and Sound (+13.37 points). These  
 503 results demonstrates that explicitly converting audio into dense, semantically rich captions—rather  
 504 than relying on latent audio tokens—provides a far more effective and universal audio representation,  
 505 successfully capturing non-lexical attributes such as musical genre, timbre, rhythm, and complex  
 506 environmental events.

507  
 508 Table 9: **Per-category breakdown of the controlled OmniEmbed ablation (R@1).** Replacing native  
 509 audio tokens with our Audio-as-Text mechanism yields large gains **across all acoustic domains**,  
 510 including non-speech categories. Best results are highlighted in **bold**.

Target Audio Category	OmniEmbed (Original)	OmniEmbed (Ours-modified)	Absolute Gain
Human Speech	11.36	<b>50.38</b>	+39.02
Music	16.07	<b>30.00</b>	+13.93
Sound	10.75	<b>24.12</b>	+13.37

## 517 6 CONCLUSION, LIMITATIONS AND FUTURE WORK

520 In this paper, we introduced **OmniCVR**, a large-scale benchmark for *omni-composed video retrieval*  
 521 that establishes vision, audio, and text as first-class modalities. By requiring models to retrieve  
 522 target videos based on source videos and natural-language instructions—spanning vision-centric,  
 523 audio-centric, and integrated queries—OmniCVR provides a rigorous testbed for multimodal com-  
 524 positionality. Our systematic evaluation reveals that existing state-of-the-art retrievers significantly  
 525 underutilize acoustic information, struggling when retrieval hinges on non-visual transformations.  
 526 To address this, we proposed **AudioVLM2Vec**, which injects explicit audio semantics into the em-  
 527 bedding pipeline. This approach achieves state-of-the-art results and exposes the failure of current  
 528 “full-modality” systems to effectively attend to speech and environmental sound cues.

529 **Limitations.** Despite its strong performance, a primary limitation is the **inference latency** caused  
 530 by the intermediate audio-to-text transcription. While effective for semantic bridging, this step incurs  
 531 higher computational overhead than latent embedding methods. Future work will focus on optimizing  
 532 this pipeline—potentially via lightweight adaptors or distillation—to accelerate embedding generation  
 533 for real-time applications.

534 **Future Work.** Moving forward, we aim to: (i) incorporate additional modalities (e.g., depth, 3D)  
 535 for richer reasoning; (ii) develop diagnostic protocols to probe temporal grounding and cross-modal  
 536 consistency; (iii) scale to cinematic long-form videos (ranging from minutes to hours, e.g., full  
 537 movies/TV episodes) by utilizing dense cropping from complex long contexts to yield short clips  
 538 with richer temporal and semantic density, thereby introducing harder distractors and larger retrieval  
 539 pools; (iv) explore open-ended retrieval to study hallucinations; and (v) leverage OmniCVR as a  
 training resource to develop robust end-to-end omni embeddings.

540 REFERENCES  
541

542 Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan  
543 Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification  
544 benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

545 Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual  
546 speech separation. In *Interspeech 2018*, 2018.

547 Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv,  
548 Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*,  
549 2024.

550 Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand  
551 Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the  
552 IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

553 Animesh Gupta, Jay Parmar, Ishan Rajendrakumar Dave, and Mubarak Shah. From play to replay:  
554 Composed video retrieval for temporally fine-grained videos. *arXiv preprint arXiv:2506.05274*,  
555 2025.

556 Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. Egocvr: An  
557 egocentric benchmark for fine-grained composed video retrieval. pp. 1–17, 2024.

558 Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. Vlm2vec:  
559 Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth  
560 International Conference on Learning Representations*, 2025.

561 Chris Kelly, Luhui Hu, Bang Yang, Yu Tian, Deshun Yang, Cindy Yang, Zaoshan Huang, Zihao Li,  
562 Jiayin Hu, and Yuexian Zou. Visiongpt: Vision-language understanding agent using generalized  
563 multimodal framework. *arXiv preprint arXiv:2403.09027*, 2024.

564 Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Hannah Recknor, Jimena  
565 Guallar-Blasco, Alexander Martin, Eugene Yang, and Benjamin Van Durme. Multivent 2.0: A  
566 massive multilingual benchmark for event-centric video retrieval. In *Proceedings of the Computer  
567 Vision and Pattern Recognition Conference*, pp. 24149–24158, 2025.

568 Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-  
569 training for unified vision-language understanding and generation. In *International conference on  
570 machine learning*, pp. 12888–12900. PMLR, 2022.

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

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

576 Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang.  
577 Valor: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions  
578 on Pattern Analysis and Machine Intelligence*, 2024.

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

583 Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef  
584 Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated  
585 video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp.  
586 2630–2640, 2019.

587 Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and  
588 Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International  
589 Conference on Learning Representations*, 2024.

594 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,  
 595 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual  
 596 models from natural language supervision. In *International conference on machine learning*, pp.  
 597 8748–8763. PMLR, 2021.

598 Gunnar A Sigurdsson, Gü̈l Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta.  
 599 Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European  
 600 conference on computer vision*, pp. 510–526. Springer, 2016.

602 Omkar Thawakar, Muzammal Naseer, Rao Muhammad Anwer, Salman Khan, Michael Felsberg,  
 603 Mubarak Shah, and Fahad Shahbaz Khan. Composed video retrieval via enriched context and  
 604 discriminative embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and  
 605 Pattern Recognition*, pp. 26896–26906, 2024.

606 Omkar Thawakar, Dmitry Demidov, Ritesh Thawakar, Rao Muhammad Anwer, Mubarak Shah,  
 607 Fahad Shahbaz Khan, and Salman Khan. Beyond simple edits: Composed video retrieval with  
 608 dense modifications. *arXiv preprint arXiv:2508.14039*, 2025.

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

613 Yuan Tseng, Layne Berry, Yi-Ting Chen, I-Hsiang Chiu, Hsuan-Hao Lin, Max Liu, Puyuan Peng,  
 614 Yi-Jen Shih, Hung-Yu Wang, Haibin Wu, et al. Av-superb: A multi-task evaluation benchmark  
 615 for audio-visual representation models. In *ICASSP 2024-2024 IEEE international conference on  
 616 acoustics, speech and signal processing (ICASSP)*, pp. 6890–6894. IEEE, 2024.

617 Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gü̈l Varol. Covr-2: Automatic data construction  
 618 for composed video retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,  
 619 2024.

621 Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu,  
 622 Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the  
 623 world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

628 Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging  
 629 video and language. In *Proceedings of the IEEE conference on computer vision and pattern  
 630 recognition*, pp. 5288–5296, 2016.

631 WU Yue, Zhaobo Qi, Yiling Wu, Junshu Sun, Yaowei Wang, and Shuhui Wang. Learning fine-  
 632 grained representations through textual token disentanglement in composed video retrieval. In *The  
 633 Thirteenth International Conference on Learning Representations*, volume 3, pp. 4, 2025.

635 Jiaqi Samantha Zhan, Crystina Zhang, Shengyao Zhuang, Xueguang Ma, and Jimmy Lin. Magmar  
 636 shared task system description: Video retrieval with omniembed. *arXiv preprint arXiv:2506.09409*,  
 637 2025.

638 Yufei Zhan, Hongyin Zhao, Yousong Zhu, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon-g:  
 639 Bridging vision-language and vision-centric tasks via large multimodal models. *arXiv preprint  
 640 arXiv:2410.16163*, 2024.

642 Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang,  
 643 Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In  
 644 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13691–13701, 2025.

645 Luwei Zhou, Chen Xu, and Jason J Corso. Towards automatic learning of procedures from web  
 646 instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32,  
 647 2018.

648 A DATASETS USED IN STAGE 1 (VIDEO CURATION & SEGMENTATION)  
649650 In Stage 1, we curate and segment long-form videos (typically minutes to hours in duration) from six  
651 complementary datasets that cover diverse domains, modalities, and annotation styles. From each of  
652 these long-form videos, we extract multiple semantically coherent short clips (5–15 seconds) that  
653 constitute the final retrieval items in our benchmark (average length 11.8 seconds). Below we provide  
654 more detailed descriptions of the six datasets: HowTo100M (Miech et al., 2019), MSR-VTT (Xu  
655 et al., 2016), VATEX (Wang et al., 2019), YouTube8m-MusicTextClips (Abu-El-Haija et al., 2016),  
656 YouCook2 (Zhou et al., 2018), and VALOR (Liu et al., 2024).  
657658 **HowTo100M (Miech et al., 2019).** A large-scale collection of narrated instructional videos from  
659 YouTube. It contains about 1.2M long videos and roughly 136M automatically transcribed narration  
660 clips, covering over 23k diverse tasks (cooking, DIY, fitness, crafts, etc.). Strengths include un-  
661 matched scale and domain coverage, making it ideal for weakly supervised pretraining of text–video  
662 embeddings. However, ASR transcripts are noisy and loosely aligned with visual content, and activity  
663 distribution is highly imbalanced. In Stage 1, it serves as the primary source of long videos, where  
664 narration timestamps provide coarse cues for segmentation.  
665666 **MSR-VTT (Xu et al., 2016).** A widely used benchmark of general-domain short clips with human-  
667 written captions. It contains 10k clips and 200k captions (about 20 per clip). Strengths are high-quality  
668 human annotations and balanced coverage across diverse scenarios. Limitations include its clip-level  
669 scope and lack of fine-grained temporal structure. In Stage 1, it is primarily used as a reference  
670 benchmark to evaluate the semantic quality of curated segments after segmentation.  
671672 **VATEX (Wang et al., 2019).** A multilingual video–caption dataset with both English and Chinese  
673 annotations. It consists of 41k videos paired with 825k captions, including 206k English–Chinese  
674 parallel sentences. The dataset enables multilingual retrieval and cross-lingual transfer. Its limitations  
675 lie in its clip-level nature and absence of dense temporal supervision. In Stage 1, it complements other  
676 datasets by providing multilingual benchmarks to test cross-lingual robustness of curated segments.  
677678 **YouTube8m-MusicTextClips (Abu-El-Haija et al., 2016).** Derived from the large-scale YouTube-  
679 8M dataset, this subset focuses on music and music-video content aligned with text or tags. YouTube-  
680 8M contains millions of videos and billions of frames with noisy machine-generated labels across  
681 4.8k entity categories. Strengths are its massive scale and coverage, especially in entertainment and  
682 music. Weaknesses are label noise and lack of dense natural-language captions. In Stage 1, it acts as  
683 a supplementary pool and weakly labeled filtering resource for music/entertainment domains.  
684685 **YouCook2 (Zhou et al., 2018).** A domain-specific dataset of cooking videos with step-level  
686 annotations. It contains 2k untrimmed YouTube videos (about 176 hours, across 89 recipes), each  
687 segmented into procedure steps with textual descriptions. It is a gold-standard benchmark for  
688 procedure segmentation and dense video–text alignment. Limitations include its domain restriction  
689 to cooking and moderate video length. In Stage 1, it is used as a calibration set for validating  
690 segmentation quality and procedure-aware alignment.  
691692 **VALOR (Liu et al., 2024).** A tri-modal dataset (vision–audio–language) designed for audiovisual  
693 captioning and retrieval. VALOR-1M provides about 1M audiovisual clips for pretraining, and  
694 VALOR-32k offers a smaller high-quality evaluation benchmark with human-curated audiovisual  
695 captions. Strengths include explicit modeling of audio cues in addition to visual and textual context,  
696 enabling tri-modal learning. Limitations include smaller scale compared to HowTo100M and potential  
697 annotation subjectivity. In Stage 1, VALOR supports audio-informed segmentation (e.g., boundary  
698 detection from speech or sound transitions) and tri-modal retrieval evaluation.  
699700 B BASELINE MODELS  
701702 To contextualize the performance of our approach, we benchmark against a diverse set of baselines,  
703 spanning both large multimodal embedding models and lightweight or task-specific retrievers. Below  
704 we provide detailed descriptions of each baseline model.  
705

702 **CLIP (Radford et al., 2021).** A foundational vision–language model trained on large-scale im-  
 703 age–text pairs. For video retrieval tasks, we follow standard practice by uniformly sampling 15  
 704 frames per video and averaging their frame-level embeddings to obtain the video representation.  
 705 While CLIP provides a strong baseline for vision–text alignment, it lacks explicit modeling of audio  
 706 or video-specific temporal dynamics.

708 **CoVR (Thawakar et al., 2024).** A model specifically designed for composed video retrieval,  
 709 focusing on scenarios where the query involves modifications of existing video content. CoVR is  
 710 optimized for vision-centric transformations, and in our setting, we uniformly sample 15 frames per  
 711 video and average their embeddings to form the video representation. We assess CoVR’s ability to  
 712 generalize beyond purely visual changes. Its lightweight design makes it efficient, though its limited  
 713 multimodal scope is a constraint for audio-aware retrieval.

715 **BLIP (Li et al., 2022).** A unified vision–language understanding and generation model. We adapt  
 716 BLIP for retrieval by using its vision–language matching head to score candidate videos. Following  
 717 our experimental protocol, we uniformly sample 15 frames per video and average their embeddings to  
 718 construct the video representation. BLIP demonstrates strong cross-modal reasoning and captioning  
 719 ability, making it a competitive retrieval baseline. However, like CLIP, it does not natively incorporate  
 720 audio cues.

721 **ImageBind (Girdhar et al., 2023).** A multimodal embedding model that learns a joint represen-  
 722 tation space across six modalities: images, text, audio, video, depth, and IMU signals. For video  
 723 retrieval, we apply the same protocol as other lightweight baselines, uniformly sampling 15 frames  
 724 per video and averaging their frame-level embeddings. Its broad modality coverage makes it a natural  
 725 fit for audio–video retrieval tasks. The ability to align heterogeneous modalities directly in a shared  
 726 space provides a strong baseline for multimodal integration, though its representations may be less  
 727 specialized than task-specific models.

729 **OmniEmbed-v0.1-multivent (Ma et al., 2025).** A unified multimodal embedding model built  
 730 on the Tevatron 2.0 framework. It is trained to encode text, image, audio, and video into a shared  
 731 representation space. OmniEmbed-v0.1-multivent has achieved state-of-the-art performance in  
 732 cross-modal video retrieval benchmarks, such as the MAGMaR Shared Task (Zhan et al., 2025), by  
 733 fine-tuning on the MultiVENT dataset (Kriz et al., 2025) with joint vision–audio–text supervision. Its  
 734 strength lies in versatility across modalities, making it a strong baseline for multimodal retrieval.

736 **VLM2Vec (Jiang et al., 2025).** A general framework that transforms existing large vision–language  
 737 models (VLMs)—including Qwen2-VL (Wang et al., 2024), LLaVA (Liu et al., 2023), and Phi-  
 738 3.5-V—into universal embedding models through contrastive learning on the Massive Multimodal  
 739 Embedding Benchmark (MMEB). VLM2Vec consistently surpasses conventional baselines such as  
 740 CLIP (Radford et al., 2021) and BLIP (Li et al., 2022), with improvements on retrieval tasks. This  
 741 approach highlights the effectiveness of adapting powerful pretrained VLMs into embedding-focused  
 742 architectures.

743 **AudioVLM2Vec (Ours).** Our proposed extension of VLM2Vec that explicitly integrates audio  
 744 semantics. We employ Qwen2-Audio (Chu et al., 2024) to generate fine-grained captions from the  
 745 video’s audio track, which are then concatenated with the user’s modification query and fed into  
 746 VLM2Vec. This design leverages the strong representation capabilities of large VLMs while ensuring  
 747 sensitivity to audio signals, thereby improving performance on audio-centric retrieval scenarios.

## 750 C SUPPLEMENTARY EXPERIMENTS AND DISCUSSION

### 753 C.1 OMNICVR DATA STATISTICS AND ANALYSIS

755 **Figure 6** shows the distribution of video lengths in the OmniCVR training dataset, where most clips  
 fall between 5 and 15 seconds.

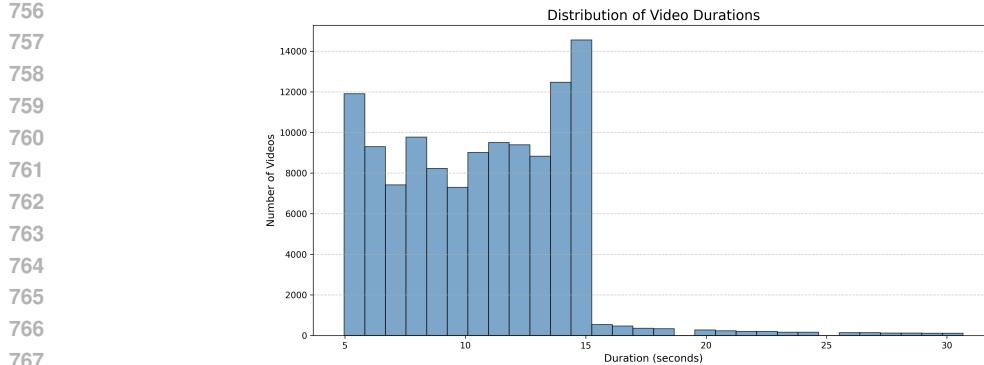


Figure 6: Video length distribution of the OmniCVR training dataset.

To better illustrate the breadth of real-world scenarios covered by OmniCVR, we provide a detailed breakdown of its content sources. OmniCVR aggregates videos from diverse publicly available datasets spanning multiple genres:

- **Instructional & Procedural Content:** Videos depicting step-by-step tasks such as cooking recipes, DIY repairs, crafting tutorials, and fitness routines. These are critical for evaluating a model’s ability to understand fine-grained actions and temporal dependencies.
- **Daily Life, Nature & Travel:** Unscripted, in-the-wild footage capturing everyday human activities, natural environments, and travel vlogs. This category provides rich visual and auditory context and serves as the primary source for environmental sound-related queries.
- **Music & Performance:** Clips featuring musical instruments, dance performances, and live concerts. These are specifically leveraged for “Audio-Centric” queries, enabling evaluation of a model’s capacity to distinguish between musical genres, tempos, instruments, and performance styles.
- **General Entertainment & Events:** A broad collection of web-sourced clips including sports highlights, news broadcasts, public ceremonies, and other event-driven content, ensuring coverage of dynamic and socially relevant scenarios.

This multi-genre composition ensures that OmniCVR comprehensively reflects the heterogeneity of real-world audiovisual experiences, making it a robust benchmark for cross-modal retrieval and understanding.

Complementing these visual genres, OmniCVR explicitly models the acoustic dimension with high granularity. To ensure comprehensive coverage of acoustic, semantic, and para-linguistic features, we categorize audio content into three primary domains, each enforcing a strict schema of fine-grained attributes:

- **Human Speech:** This category addresses both the *lexical* and *para-linguistic* dimensions of spoken audio. Our annotation pipeline explicitly captures:
  - *Lexical Content:* Verbatim transcripts of the speech to ground accurate semantic understanding.
  - *Para-linguistic Features:* Identification of speaker characteristics and emotional tone (e.g., neutral, angry, fearful, surprised), distinguishing retrieval targets based on *how* something is said, not just *what* is said.
- **Music:** This category focuses on *temporal dynamics* and stylistic attributes. The annotations provide detailed descriptions of:
  - *Genre & Instrumentation:* Identification of specific musical styles and the instruments present.
  - *Temporal Dynamics:* Chronological descriptions of rhythm, pace, and intensity (e.g., distinguishing a “slow, steady beat” from a “fast, erratic tempo”) and the overall atmospheric mood.

810  
 811 • **Environmental Sound**: To address the complexity of acoustic environments, we enforce a  
 812 detailed *hierarchy* distinguishing between:  
 813  
 814 – *Nature Sounds*: Elements such as wind, rain, water flow, and animal calls.  
 815 – *Mechanical & Urban Soundscapes*: Sounds of engines, machinery, alarms, traffic  
 816 patterns, and construction noise.  
 817 – *Foley & Action Sounds*: Distinct, event-driven sounds such as footsteps, glass breaking,  
 818 or doors closing.  
 819

820 By explicitly modeling these dimensions during the annotation and verification stages, OmniCVR  
 821 ensures that retrieval queries are grounded in rich, fine-grained audio semantics rather than simplified  
 822 labels.  
 823  
 824

## 825 C.2 VISION-CENTRIC RETRIEVAL RESULTS

827 Table 10 reports vision-centric retrieval results on OmniCVR. Within lightweight baselines, IMAGE-  
 828 BIND achieves the best performance, indicating that simple cross-modal alignment still provides  
 829 competitive vision retrieval ability. Yet, large multimodal embedding models clearly dominate:  
 830 VLM2Vec already surpasses OmniEmbed-v0.1-multivent, and our AudioVLM2Vec further advances  
 831 state-of-the-art performance by a large margin (e.g., +14.3% R@1 over VLM2Vec). Interestingly,  
 832 the consistent gains obtained by AudioVLM2Vec—even under vision-only evaluation—suggest that  
 833 audio representations implicitly encode complementary visual characteristics, which can be leveraged  
 834 during joint embedding learning. This finding highlights the underexplored synergy between audio  
 835 and vision modalities: audio not only enriches multimodal fusion but also strengthens pure visual  
 836 retrieval through cross-modal feature transfer.  
 837

838 Table 10: Vision-centric retrieval performance on OmniCVR. We report Recall at K (R@1, R@3,  
 839 R@5, R@10). Best results within each category are highlighted in **bold**.  
 840

841 <b>Models</b>	842 <b>Backbone</b>	843 R@1	844 R@3	845 R@5	846 R@10
<b>Lightweight and Task-Specific Models</b>					
844 CLIP	845 CLIP	39.53	51.18	56.27	63.10
845 CoVR	846 BLIP2	2.10	6.57	9.29	14.46
846 BLIP	847 BLIP	6.13	8.85	10.60	13.41
847 IMAGEBIND	CLIP	<b>47.85</b>	<b>57.58</b>	<b>61.09</b>	<b>65.82</b>
<b>Large Multimodal Embedding Models</b>					
848 OmniEmbed-v0.1-multivent	849 Qwen2.5-Omni	50.74	63.45	69.85	77.21
849 VLM2Vec	850 Qwen2-VL	55.04	66.78	70.99	75.99
850 AudioVLM2Vec (Ours)	851 Qwen2-Audio + Qwen2-VL	<b>69.33</b>	<b>76.51</b>	<b>80.28</b>	<b>82.82</b>

## 853 C.3 GENERALIZATION TO OUT-OF-DOMAIN TASKS

856 To further validate the generalization capability of the proposed audio-augmented representations, we  
 857 evaluate AudioVLM2Vec in a zero-shot manner on the widely used MSR-VTT dataset (Xu et al.,  
 858 2016) for conventional text-to-video retrieval. As shown in Table 11, AudioVLM2Vec consistently  
 859 outperforms the strong VLM2Vec baseline across all metrics (+1.7–2.3 absolute points), even though  
 860 MSR-VTT captions rarely mention sound events explicitly. Qualitative analysis reveals that our audio-  
 861 to-text descriptions provide valuable implicit visual context (e.g., “roaring engines and screeching  
 862 tires” in car racing clips, “crowd cheering” in sports scenes), thereby enhancing discrimination  
 863 among visually similar videos. This result confirms that the multimodal representations learned on  
 864 OmniCVR transfer effectively to standard retrieval tasks.

864  
 865  
 866  
 867  
 868 Table 11: Zero-shot text-to-video retrieval performance on MSR-VTT. Our AudioVLM2Vec is trained  
 869 only on OmniCVR and evaluated using its automatically generated audio-to-text descriptions as  
 870 additional input. Best results are highlighted in **bold**.  
 871  
 872

Model	R@1	R@3	R@5	R@10
VLM2Vec	36.10	53.00	60.70	70.10
AudioVLM2Vec (Ours)	<b>37.90</b>	<b>55.30</b>	<b>62.50</b>	<b>71.80</b>
Δ	<b>+1.80</b>	<b>+2.30</b>	<b>+1.80</b>	<b>+1.70</b>

873 C.4 DISCUSSION ON THE PLAUSIBILITY OF COMPOSITIONAL TRIPLETS  
 874

875 We clarify that our benchmark construction is strictly data-driven. We do not synthesize arbitrary  
 876 instructions; instead, we first mine valid video pairs (Source, Target) from real-world distributions  
 877 (e.g., HowTo100M, YouTube8M) based on semantic similarity. The modification text is generated  
 878 post-hoc to describe the actual physical and acoustic differences observed between the clips. Thus, all  
 879 “transformations” in OmniCVR reflect natural variations found in large-scale video corpora, ensuring  
 880 that matching targets physically exist within the dataset.  
 881

882 C.5 WHY DO “FULL-MODALITY” METHODS DIVERGE ON AUDIO-CENTRIC QUERIES?  
 883

884 Although ImageBind, OmniEmbed, and VLM2Vec all operate in multi-modal settings, their behaviors  
 885 differ markedly once retrieval is driven *only* by audio changes.  
 886

887 **ImageBind.** ImageBind is capable of ingesting audio and text, yet in our audio-centric OmniCVR  
 888 splits—where speech-guided queries constitute a majority—it struggles to capture the *lexical* and  
 889 *semantic* content of speech. Its audio pathway is optimized for generic cross-modal alignment rather  
 890 than fine-grained speech understanding; as a result, instructions that hinge on who spoke, what was  
 891 said, or subtle speech-state changes (e.g., tone, intent) are often mapped to embeddings dominated  
 892 by background acoustics or coarse timbral cues. This mismatch can even *mislead* retrieval under  
 893 audio-centric composition, causing the search to prioritize scenes with similar ambient sounds while  
 894 ignoring the intended speech-driven modification.  
 895

896 **OmniEmbed.** OmniEmbed can encode audio, but its token budget and fusion design weigh  
 897 visual tokens much more heavily than audio tokens across the sequence. During joint pooling, the  
 898 resultant embedding is therefore dominated by vision features, and audio contributes weakly to  
 899 the final representation. Under audio-centric queries—where the visual stream is deliberately held  
 900 constant—this imbalance suppresses precisely the information that differentiates the targets, yielding  
 901 poor retrieval.  
 902

903 **VLM2Vec.** VLM2Vec does not include an explicit audio branch. Consequently, it performs com-  
 904 petitively in vision-centric and integrated (vision+audio) settings—where the visual stream provides  
 905 reliable discriminative cues and the composed instruction contains strong visual operators—but fails  
 906 when the *only* changing factor is audio. In audio-centric queries, the absence of an audio encoder  
 907 leaves the model with no path to ground the modification.  
 908

909 **AudioVLM2Vec (ours).** In contrast, our approach leverages **Qwen2-Audio** to first convert raw  
 910 audio—especially speech—into fine-grained textual descriptions. These transcribed and summarized  
 911 audio semantics are then injected into the VLM2Vec pipeline, where they are aligned with vision-  
 912 language embeddings. This design bypasses the limitations of weak audio tokenization and ensures  
 913 that speech content is represented in the same space as visual and textual cues. As reflected by the  
 914 audio-centric results, AudioVLM2Vec achieves **77.2 R@1** and **94.2 R@10**, far beyond all other  
 915 full-modality methods, showing that preserving semantic detail from audio before fusion is a decisive  
 916 advantage.  
 917

918 In summary, the failure modes align with architectural choices: (i) speech-heavy audio-centric  
 919 composition penalizes models that lack *speech-aware* audio representations, (ii) token and pooling  
 920 imbalance can *dilute* audio contributions, and (iii) omitting an audio branch altogether leads to  
 921 systematic failure whenever sound is the only supervisory signal. By contrast, our design demonstrates  
 922 that converting audio to rich textual descriptions before multimodal alignment allows audio to function  
 923 as a true *first-class* compositional signal in OmniCVR.  
 924

918 **D ADDITIONAL QUALITATIVE EXAMPLES**  
919  
920  
921  
922  
923  
924

925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

*Keep the boy and couch setting. Remove all sickness elements (tissues, recovery dialogue). Replace with playful physical movement on the couch, ending in a cheerful mutual "goodbye." Change the visual action and conversational tone from caring to playful.*



Figure 7: Qualitative example of modifying emotional tone and action. The query instructs the model to **remove sickness elements** (tissues, recovery dialogue) and **replace them with playful physical movement**, shifting the scene from caring to cheerful while maintaining the characters and setting.



972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999

*Maintain the graffiti artist, the wall, and the Hip-Hop soundtrack genre. Advance the visual state of the wall art from an early, mostly white outline to a nearly finished, detailed piece with red and black fill.*



Figure 8: Qualitative example of temporal and state progression. The model is tasked to **advance the visual state** of the graffiti wall from an early outline to a *nearly finished, detailed piece*, while preserving the specific artist and Hip-Hop audio backdrop.



972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025

Figure 9: Qualitative example of fine-grained integrated modification. The instruction requires simultaneous changes in vision (changing shirt color from yellow to white/light gray) and audio (depicting a faster-paced sequence of throws), testing the model’s ability to handle precise attribute.



Figure 10: Qualitative example of semantic conversation shift. The query directs a transition from recollecting past events to discussing **future possibilities** (e.g., enrolling in Capoeira), requiring the model to understand the semantic content of the speech to retrieve the correct forward-looking segment.

We provide additional qualitative results to demonstrate the model’s capability in handling diverse and complex compositional instructions. **Figure 7** illustrates a vision-centric transformation of emotional tone, shifting a scene from “caring recovery” to “playful interaction.” **Figure 8** captures a temporal state progression, where the query directs the retrieval of a completed graffiti piece based on an early outline. **Figure 9** showcases a fine-grained integrated query, requiring simultaneous reasoning over visual attributes (shirt color change) and auditory events (pace of throws). Finally, **Figure 10** demonstrates a semantic conversation shift, where the model must distinguish between an interview segment about past recollections versus one focused on future aspirations.

## E ETHICS STATEMENT

The OmniCVR benchmark proposed in this work is derived from existing public video datasets, including HowTo100M, MSR-VTT, VATEX, YouTube8M, YouCook2, and VALOR. We strictly

1026 adhere to the respective licenses and terms of use for these source datasets. The videos originate  
 1027 from publicly available content and do not involve private information belonging to individuals or  
 1028 organizations beyond what is already public.

1029 For the human verification stage of our pipeline (Stage 3), we employed expert annotators to  
 1030 validate the quality of the modification instructions and video pairings. We ensured ethical working  
 1031 conditions and paid all annotators a fair hourly wage that exceeds the local minimum wage standards.  
 1032 Furthermore, during the generative annotation process involving Large Multimodal Models (LMMs),  
 1033 we applied safety filters to prevent the inclusion of harmful, offensive, or biased content in the  
 1034 generated text descriptions.

1035

## 1036 F REPRODUCIBILITY STATEMENT

1037 We are committed to ensuring the reproducibility of our results and promoting further research in  
 1038 omni-modal retrieval.

1039

- 1040 • **Data Availability:** The OmniCVR benchmark, including the 160k+ curated clips, the 50k+  
 1041 compositional triplets, and the gold-standard test set, will be made fully open-source to the  
 1042 public upon publication.
- 1043 • **Code and Models:** Our proposed AudioVLM2Vec framework leverages open-source pre-  
 1044 trained weights (Qwen2-Audio and Qwen2-VL) and standard libraries. We will release  
 1045 the complete codebase, including data generation scripts, training code, and evaluation  
 1046 protocols.
- 1047 • **Transparency:** To facilitate the reproduction of our dataset construction pipeline, we  
 1048 have provided the exact prompts used for annotation, triplet mining, and verification in  
 1049 Appendix G.

1050

1051

## 1052 G PROMPTS

1053

1054 Below are the prompts used for triplet mining.

1055

1056

1057

### Prompt for generating chronological video captions using Qwen2.5-Omni

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

You are a precise observer. Write one paragraph that describes ONLY what is directly visible  
 and audible in the video, in strict chronological order with clear temporal markers.

For any notable action, break it into distinct stages and describe each in detail (e.g., starting  
 slowly, changing technique, altering body position, adding props). If a stage is missing or  
 unclear, state “not shown” or “unclear.” This is the highest priority. Do not just state an  
 activity; describe precisely how it unfolds and changes over time. Detail the sequence of  
 movements. Example: Instead of “A person plays football,” you must write: “A person begins  
 by slowly dribbling a white and black ball with their right foot across a green field. They  
 then transition to kicking it against a wall, and later, after lying on their back, they attempt to  
 juggle the ball with their feet in the air.”

#### Rules:

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

- No guesses or world knowledge.
- If something is uncertain, say “unclear”.
- Include at least one sentence about **AUDIO** (speech, sound, or music). If silent,  
 state “no audible speech; ambient silence/noise”. If there were voices, specify who  
 spoke, what was said in the video, and the emotion conveyed; if background music  
 (BGM) is present, describe the genre and mood of the BGM; if there are other  
 sounds, indicate their nature.
- Prefer concrete attributes (colors, materials, relative positions) over interpretations.

1080  
1081  
1082  
10831084 Prompt for generating vision-only video modification instructions  
1085

1086 You are an expert in vision-only video understanding and creative language. Given two  
 1087 textual descriptions—[Source Description] and [Target Description]—generate a concise,  
 1088 natural-language instruction that tells someone how to modify the visual content of the  
 1089 “Source Video” to match the “Target Video”.

1090 Your instruction must:

- 1091 • Focus only on **visible differences** (e.g., added/removed objects, changed actions,  
 1092 different person, altered background, lighting, color tone, camera view, or scene  
 1093 layout).
- 1094 • Completely ignore sound, audio, dialogue, music, or any non-visual information.
- 1095 • Be phrased as a clear, direct command or user request.
- 1096 • Start from the visual context of the Source Video.

1098 Now, generate a vision-only instruction for the following descriptions:  
1099

1100 [Source Description]: {source\_desc}

1101 [Target Description]: {target\_desc}

1102 **Generated Instruction:**

1103  
1104  
1105  
1106  
1107  
1108  
1109  
11101111 Prompt for generating audio-only video modification instructions  
1112

1113 You are an expert in audio-only video understanding and creative language. Given two textual  
 1114 descriptions—[Source Description] and [Target Description]—generate a concise, natural-  
 1115 language instruction that tells someone how to modify the audio content of the “Source Video”  
 1116 to match the “Target Video”.

1117 Your instruction must:

- 1118 • Focus only on **audible differences** (e.g., added/removed speech, changed speaker,  
 1119 background sounds, sound effects, music type, volume, or tone).
- 1120 • Completely ignore any visual information (objects, actions, people, backgrounds,  
 1121 colors, lighting, or camera view).
- 1122 • Be phrased as a clear, direct command or user request.
- 1123 • Start from the audio context of the Source Video.

1126 Now, generate an audio-only instruction for the following descriptions:  
1127

1128 [Source Description]: {source\_desc}

1129 [Target Description]: {target\_desc}

1130 **Generated Instruction:**

1131  
1132  
1133

1134

1135

1136

1137

## Prompt for generating video-to-video modification instructions

1138

1139

1140

1141

You are an expert in creative language and video retrieval. Your task is to generate a concise, natural language instruction that describes how to modify a “Source Video” to become a “Target Video”. You will be given two textual descriptions: [Source Description] and [Target Description].

1142

Your instruction should:

1143

1144

1145

1146

1147

1148

1149

- Identify the most salient difference(s) between the source and target (e.g., scene, action, object, person, setting).
- Be phrased as a clear command or user request.
- Be natural and easy to understand.
- Start from the context of the Source Video.

1150

**Now, generate an instruction for the following descriptions:**

[Source Description]: {source\_desc}

[Target Description]: {target\_desc}

**Generated Instruction:**

1151

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

## Prompt for generating detailed audio descriptions

1162

1163

You are an expert audio analyst. Listen carefully to the given audio and provide a comprehensive description.

1164

**Instructions:**

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

- First, identify the **main category** of the audio. Choose exactly one from: human\\_speech, music, environmental\\_sound, event\\_sound, other.
- Then, provide a **detailed explanation** according to the detected category:
  - **Human speech:** Transcribe the speech verbatim. Then identify the speaker’s emotion (choose from: neutral, happy, sad, angry, fearful, surprised, disgusted).
  - **Music:** Describe genre, instruments, vocals (if any), rhythm, and the overall mood.
  - **Environmental sound:** Describe the environment, natural elements, and the acoustic atmosphere.
  - **Event sound:** Describe the specific event or action represented by the sound, including temporal sequence if clear.
  - **Other:** Provide as precise a description as possible of what is heard.
- Finally, summarize the entire audio in **one concise sentence**.

**Rules:**

1180

1181

1182

1183

1184

1185

1186

1187