# Novelty Assessment Report

**Paper**: WorldSense: Evaluating Real-world Omnimodal Understanding for Multimodal LLMs
**PDF URL**: https://openreview.net/pdf?id=YxsfxAvJv4
**Venue**: ICLR 2026 Conference Submission
**Year**: 2026
**Report Generated**: 2026-01-01

## Abstract

We introduce WorldSense, the first benchmark to assess the multi-modal video understanding, that simultaneously encompasses visual, audio, and text inputs. In contrast to existing benchmarks, our WorldSense has several features: (i) collaboration of omni-modality, we design the evaluation tasks to feature a strong coupling of audio and video, requiring models to effectively utilize the synergistic perception of omni-modality; (ii) diversity of videos and tasks, WorldSense encompasses a diverse collection of 1,662 audio-visual synchronised videos, systematically categorized into 8 primary domains and 67 fine-grained subcategories to cover the broad scenarios, and 3,172 multi-choice QA pairs across 26 distinct tasks to enable the comprehensive evaluation; (iii) high-quality annotations, all the QA pairs are manually labeled by 80 expert annotators with multiple rounds of correction to ensure quality. Based on our WorldSense, we extensively evaluate various state-of-the-art models. The experimental results indicate that existing models face significant challenges in understanding real-world scenarios (65.1% best accuracy). By analyzing the limitations of current models, we aim to provide valuable insight to guide development of real-world understanding. We hope our WorldSense can provide a platform for evaluating the ability in constructing and understanding coherent contexts from omni-modality.



## Core Task Landscape

This paper addresses: **Omnimodal Video Understanding with Synchronized Audio and Visual Inputs**

A total of **50 papers** were analyzed and organized into a taxonomy with **21 categories**.

### Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Multimodal Representation Learning and Alignment**
- **Omnimodal Large Language Models and Interactive Systems**
- **Audio-Visual Generation with Temporal Control**
- **Task-Specific Audio-Visual Understanding**
- **Benchmarks and Evaluation Frameworks**
- **Multimodal Analysis and Theoretical Frameworks**

### Complete Taxonomy Tree

- Omnimodal Video Understanding with Synchronized Audio and Visual Inputs Survey Taxonomy
- Multimodal Representation Learning and Alignment
  - Tri-modal Pre-training and Contrastive Learning (3 papers)
  - [1] Unified video-language pre-training with synchronized audio (Mo, 2024) View paper
  - [6] Representation learning for semantic alignment of language, audio, and visual modalities (Sudarsanam Parthasaarathy, 2025) View paper
  - [19] Quality over quantity? llm-based curation for a data-efficient audio-video foundation model (Vosoughi, 2025) View paper
  - Audio-Visual Synchronization and Temporal Alignment (3 papers)
  - [8] Audio-visual scene analysis with self-supervised multisensory features (Andrew Owens, 2018) View paper
  - [10] Intra-modal and Cross-modal Synchronization for Audio-visual Deepfake Detection and Temporal Localization (A Anshul, 2025) View paper
  - [22] SVLTA: Benchmarking Vision-Language Temporal Alignment via Synthetic Video Situation (Hao Du, 2025) View paper
  - Cross-modal Attention and Fusion Mechanisms (3 papers)
  - [3] Multimodal alignment and fusion: A survey (Li Songtao, 2024) View paper
  - [4] Cross-modal prompts: Adapting large pre-trained models for audio-visual downstream tasks (Duan, 2023) View paper
  - [15] Ma-avt: Modality alignment for parameter-efficient audio-visual transformers (Tanvir Mahmud, 2024) View paper
  - Multimodal Cognitive and Dual-System Architectures (2 papers)
  - [25] Omni-R1: Reinforcement Learning for Omnimodal Reasoning via Two-System Collaboration (Zhong Hao, 2025) View paper
  - [35] Coavt: A cognition-inspired unified audio-visual-text pre-training model for multimodal processing (Xianghu Yue, 2025) View paper
- Omnimodal Large Language Models and Interactive Systems
  - Real-time Streaming and Interactive Dialogue Systems (4 papers)
  - [11] X-Streamer: Unified Human World Modeling with Audiovisual Interaction (Xie You, 2025) View paper
  - [27] Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue (Tong Wenwen, 2025) View paper
  - [36] Internlm-xcomposer2. 5-omnilive: A comprehensive multimodal system for long-term streaming video and audio interactions (Zhang Pan, 2024) View paper
  - [44] LongCat-Flash-Omni Technical Report (Meituan LongCat Team, 2025) View paper
  - Unified Omnimodal Understanding and Generation (2 papers)

    - [20] Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions (Zhang Guozhen, 2025) View paper
    - [32] JavisGPT: A Unified Multi-modal LLM for Sounding-Video Comprehension and Generation (Kai Liu, 2025) View paper
    - Audio-Visual Question Answering with LLMs (3 papers)
    - [9] Watch and Listen: Understanding Audio-Visual-Speech Moments with Multimodal LLM (Li, 2025) View paper
    - [38] CLIP-Powered TASS: Target-Aware Single-Stream Network for Audio-Visual Question Answering (Yuanyuan Jiang, 2025) View paper
    - [39] Answering Diverse Questions via Text Attached with Key Audio-Visual Clues (Yu, 2024) View paper
- Audio-Visual Generation with Temporal Control
    - Video-to-Audio Synthesis (3 papers)
    - [18] Long-Video Audio Synthesis with Multi-Agent Collaboration (Xu Xinli, 2025) View paper
    - [24] Controllable Video-to-Music Generation with Multiple Time-Varying Conditions (Junxian Wu, 2025) View paper
    - [45] Video-to-audio generation with hidden alignment (Xu, 2024) View paper
    - Audio-to-Video Synthesis (2 papers)
    - [13] Audio-Sync Video Generation with Multi-Stream Temporal Control (Weng, 2025) View paper
    - [23] Diverse and aligned audio-to-video generation via text-to-video model adaptation (Itai Gat, 2024) View paper
    - Joint Audio-Video Generation (3 papers)
    - [14] Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization (Liu Kai, 2025) View paper
    - [16] Av-link: Temporally-aligned diffusion features for cross-modal audio-video generation (Menapace, 2025) View paper
    - [40] Aligning What Matters: Masked Latent Adaptation for Text-to-Audio-Video Generation (J Zheng, 2025) View paper
- Task-Specific Audio-Visual Understanding
    - Audio-Visual Segmentation and Localization (3 papers)
    - [21] Robust Audio-Visual Segmentation via Audio-Guided Visual Convergent Alignment (Liu Chen, 2025) View paper
    - [29] Unified multisensory perception: Weakly-supervised audio-visual video parsing (Yapeng Tian, 2020) View paper
    - [47] AVS-Mamba: Exploring Temporal and Multi-Modal Mamba for Audio-Visual Segmentation (Zhang Lu, 2025) View paper
    - Temporal Event Understanding and Moment Retrieval (2 papers)
    - [30] 2DP-2MRC: 2-Dimensional Pointer-based Machine Reading Comprehension Method for Multimodal Moment Retrieval (Jiajun He, 2024) View paper
    - [46] Interactive Retrieval System for Multi-Stream Collections: multiXview at CASTLE 2025 Interactive Grand Challenge (Omar Shahbaz Khan, 2025) View paper
    - Audio-Visual Scene-Aware Dialogue (2 papers)
    - [26] Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog (Zekang Li, 2021) View paper
    - [43] Enhancing Cross-Modal Understanding for Audio Visual Scene-Aware Dialog Through Contrastive Learning (Feifei Xu, 2024) View paper
    - Visual Speech Recognition and Lip Synchronization (2 papers)
    - [31] AlignVSR: Audio-visual cross-modal alignment for visual speech recognition (Liu ZeHua, 2024) View paper
    - [41] A Bridge from Audio to Video: Phoneme-Viseme Alignment Allows Every Face to Speak Multiple Languages (Wei Kun, 2025) View paper
    - Active Speaker Detection (1 papers)
    - [48] AS-Net: active speaker detection using deep audio-visual attention (Abduljalil Radman, 2024) View paper
    - Multilingual Visual Answer Localization (1 papers)
    - [50] Learning to unify audio, visual and text for audio-enhanced multilingual visual answer localization (Wen Zhibin, 2024) View paper
- Benchmarks and Evaluation Frameworks
    - Omnimodal Comprehension Benchmarks ★ (4 papers)
    - [0] WorldSense: Evaluating Real-world Omnimodal Understanding for Multimodal LLMs (Anon et al., 2026) View paper
    - [5] OmniEval: A Benchmark for Evaluating Omni-modal Models with Visual, Auditory, and Textual Inputs (Zhang Yiman, 2025) View paper
    - [12] Daily-Omni: Towards Audio-Visual Reasoning with Temporal Alignment across Modalities (Zhou Ziwei, 2025) View paper
    - [17] OmniMMI: A Comprehensive Multi-modal Interaction Benchmark in Streaming Video Contexts (Wang Yu-xuan, 2025) View paper
    - Audio-Centric and Audio-Visual Evaluation (2 papers)
    - [2] Audio-centric video understanding benchmark without text shortcut (Yudong Yang, 2025) View paper
    - [28] Vggsounder: Audio-visual evaluations for foundation models (Wiedemer, 2025) View paper
    - Long-form and Streaming Video Benchmarks (2 papers)
    - [33] Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos (Tiantian Geng, 2025) View paper
    - [37] AVoCaDO: An Audiovisual Video Captioner Driven by Temporal Orchestration (Chen, 2025) View paper
    - Specialized Domain and Neuroscience Datasets (3 papers)
    - [7] CineBrain: A Large-Scale Multi-Modal Brain Dataset During Naturalistic Audiovisual Narrative Processing (Gao Jian-xiong, 2025) View paper
    - [42] Harmonizing Sight and Sound: The Impact of Auditory Emotional Arousal, Visual Variation, and Their Congruence on Consumer Engagement in Short Video Marketing (Qiang Yang, 2025) View paper
    - [49] Multi-dimensional fusion: transformer and GANs-based multimodal audiovisual perception robot for musical performance art (Shiyi Lu, 2023) View paper
- Multimodal Analysis and Theoretical Frameworks (1 papers)
    - [34] Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis (Prathusha Sarma, 2020) View paper

## Narrative

Core task: omnimodal video understanding with synchronized audio and visual inputs. The field has evolved into several major branches that reflect different emphases on representation, interaction, generation, and evaluation. Multimodal Representation Learning and Alignment focuses on learning joint embeddings and cross-modal correspondences, often drawing on contrastive or canonical correlation techniques to align audio and visual streams. Omnimodal Large Language Models and Interactive Systems integrate these modalities into conversational agents and reasoning frameworks, enabling richer human-machine interaction. Audio-Visual Generation with Temporal

Control addresses synthesis tasks where temporal synchronization is paramount, while Task-Specific Audio-Visual Understanding targets specialized problems such as sound source localization or audio-visual scene analysis. Benchmarks and Evaluation Frameworks provide standardized testbeds for measuring omnimodal comprehension, and Multimodal Analysis and Theoretical Frameworks explore the underlying principles that govern cross-modal fusion. Representative works like Unified Video Language Audio[1] and Audio Centric Video[2] illustrate how different branches tackle alignment and task design, while foundational studies such as Audio Visual Scene Analysis[8] laid early groundwork for the field.

Recent activity highlights a tension between holistic omnimodal reasoning and task-specific optimization. Many studies pursue end-to-end architectures that unify audio, vision, and language, yet specialized benchmarks reveal that general-purpose models often struggle with fine-grained temporal or semantic alignment. WorldSense[0] sits within the Benchmarks and Evaluation Frameworks branch, specifically under Omnimodal Comprehension Benchmarks, alongside works like OmniEval[5], Daily-Omni[12], and OmniMMI[17]. While OmniEval[5] and Daily-Omni[12] emphasize diverse question types and everyday scenarios, WorldSense[0] appears to focus on comprehensive evaluation of synchronized audio-visual understanding, providing a testbed that complements these neighboring efforts. This cluster of benchmarks collectively addresses the need for rigorous assessment of omnimodal systems, ensuring that advances in representation learning and interactive models translate into measurable improvements across varied real-world tasks.

## Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

### 1. OmniEval: A Benchmark for Evaluating Omni-modal Models with Visual, Auditory, and Textual Inputs

**Authors**: Zhang Yiman, Luo Zi-heng, Yiman Zhang, Yan, Qiangyu, et al. (15 authors total) | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

In this paper, we introduce OmniEval, a benchmark for evaluating omni-modality models like MiniCPM-O 2.6, which encompasses visual, auditory, and textual inputs. Compared with existing benchmarks, our OmniEval has several distinctive features: (i) Full-modal collaboration: We design evaluation tasks that highlight the strong coupling between audio and video, requiring models to effectively leverage the collaborative perception of all modalities; (ii) Diversity of videos: OmniEval includes 810 au...

#### Relationship Analysis

Both papers belong to the Omnimodal Comprehension Benchmarks category, focusing on evaluating models' ability to understand synchronized audio-visual inputs. They overlap in their emphasis on audio-visual coupling, diverse video content, and multi-task evaluation frameworks for assessing omnimodal understanding. The key differences are that WorldSense focuses on 1,662 videos with 3,172 multiple-choice QA pairs across 26 tasks emphasizing real-world scenarios, while OmniEval includes 810 videos with 2,617 QA pairs (both open-ended and multiple-choice) across 12 sub-tasks, introducing fine-grained temporal grounding tasks and bilingual support (Chinese and English).

---

### 2. Daily-Omni: Towards Audio-Visual Reasoning with Temporal Alignment across Modalities

**Authors**: Zhou Ziwei, Wang Rui, Ziwei Zhou, Wu, Zuxuan, et al. (7 authors total) | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

Recent Multimodal Large Language Models (MLLMs) achieve promising performance on visual and audio benchmarks independently. However, the ability of these models to process cross-modal information synchronously remains largely unexplored. In this paper, we introduce: 1) Daily-Omni, an Audio-Visual Questioning and Answering benchmark comprising 684 videos of daily life scenarios from diverse sources, rich in both audio and visual information, and featuring 1197 multiple-choice QA pairs across 6 ma...

#### Relationship Analysis

Both papers belong to the Omnimodal Comprehension Benchmarks category, focusing on evaluating models' ability to understand synchronized audio-visual inputs through multiple-choice question-answering tasks. They overlap in assessing real-world scenarios requiring integrated audio-visual reasoning across diverse domains and task types. However, WorldSense emphasizes comprehensive domain coverage with 1,662 videos across 8 domains and 26 tasks with extensive manual annotation, while Daily-Omni focuses on scalability through an automated QA generation pipeline with 684 videos across 6 tasks and introduces a training-free agent baseline for temporal alignment.

---

### 3. OmniMMI: A Comprehensive Multi-modal Interaction Benchmark in Streaming Video Contexts

**Authors**: Wang Yu-xuan, Wang Yueqian, Chen Bo, Wu, Tong, et al. (8 authors total) | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

The rapid advancement of multi-modal language models (MLLMs) like GPT-4o has propelled the development of Omni language models, designed to process and proactively respond to continuous streams of multi-modal data. Despite their potential, evaluating their real-world interactive capabilities in streaming video contexts remains a formidable challenge. In this work, we introduce OmniMMI, a comprehensive multi-modal interaction benchmark tailored for OmniLLMs in streaming video contexts. OmniMMI en...

#### Relationship Analysis

Both papers belong to the Omnimodal Comprehension Benchmarks category, focusing on evaluating models' ability to understand synchronized audio-visual inputs in video contexts. They overlap in assessing multimodal video understanding with audio integration, diverse task coverage, and manual annotation quality control. However, WorldSense emphasizes real-world omnimodal understanding across 8 domains with 26 tasks requiring tight audio-visual coupling, while OmniMMI specifically targets streaming video understanding and proactive reasoning capabilities with 6 subtasks focused on interactive scenarios, multi-turn dependencies, and real-time response generation.

## Contributions Analysis

**Overall novelty summary.** WorldSense introduces a benchmark for omnimodal video understanding that requires synchronized audio-visual reasoning across 1,662 videos, 8 domains, 67 subcategories, and 3,172 QA pairs spanning 26 tasks. It resides in the 'Omnimodal Comprehension Benchmarks' leaf alongside three sibling papers: OmniEval, Daily-Omni, and OmniMMI. This leaf is part of the broader 'Benchmarks and Evaluation Frameworks' branch, which contains four distinct evaluation categories within a 50-paper taxonomy. The concentration of four papers in this specific leaf suggests a moderately active research direction focused on comprehensive omnimodal assessment.

The taxonomy reveals that WorldSense's parent branch sits alongside five other major research directions: Multimodal Representation Learning, Omnimodal LLMs, Audio-Visual Generation, Task-Specific Understanding, and Theoretical Frameworks. Neighboring evaluation categories include Audio-Centric benchmarks (2 papers), Long-form Video benchmarks (2 papers), and Specialized Domain datasets (3 papers). WorldSense's emphasis on synchronized audio-visual coupling distinguishes it from Audio-Centric benchmarks that prioritize auditory information, while its focus on diverse real-world scenarios differentiates it from Specialized Domain datasets targeting narrow applications like brain signal decoding or robotic performance art.

Among 30 candidates examined, the contribution 'Design principles emphasizing omnimodal collaboration' shows one refutable candidate from 10 examined, suggesting some overlap with prior work on audio-visual coupling requirements. The 'WorldSense benchmark' contribution itself examined 10 candidates with zero refutations, indicating potential novelty in its specific combination of scale, task diversity, and annotation quality. The 'Comprehensive evaluation revealing MLLM limitations' contribution also found no refutations among 10 candidates, though this may reflect the limited search scope rather than definitive novelty. The analysis does not cover exhaustive comparison with all existing benchmarks in the field.

Given the limited 30-candidate search scope, WorldSense appears to occupy a recognizable position within an active but not overcrowded evaluation subfield. The presence of three sibling benchmarks suggests incremental progress rather than pioneering work, yet the specific emphasis on omnimodal collaboration and the scale of manual annotation may offer distinguishing features. The analysis captures top semantic matches and immediate neighbors but cannot assess whether similar benchmarks exist outside this search radius or in adjacent research communities.

---

This paper presents **3 main contributions**, each analyzed against relevant prior work:

## Contribution 1: WorldSense benchmark for omnimodal video understanding

**Description**: The authors present WorldSense, a novel benchmark specifically designed to evaluate multimodal large language models on their ability to understand real-world scenarios through integrated processing of visual, audio, and textual information from synchronized videos. The benchmark features 1,662 videos across 8 domains and 67 subcategories, with 3,172 manually annotated question-answer pairs spanning 26 distinct tasks.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

### 1. Internvid: A large-scale video-text dataset for multimodal understanding and generation

**URL**: View paper

**Brief Assessment**

InternVid[75] focuses on large-scale video-text dataset construction for contrastive learning and generation tasks, not on creating evaluation benchmarks for omnimodal understanding with audio-visual-text integration and manually annotated QA pairs.

---

### 2. Iv-bench: A benchmark for image-grounded video perception and reasoning in multimodal llms

**URL**: View paper

**Brief Assessment**

IV-Bench[76] focuses on image-grounded video perception and reasoning with external image-text queries, whereas WorldSense emphasizes omnimodal (audio-visual-text) integration from synchronized videos. The modalities, task designs, and evaluation paradigms differ fundamentally.

---

### 3. Vidi: Large multimodal models for video understanding and editing

**URL**: View paper

**Brief Assessment**

Vidi[71] focuses on temporal retrieval tasks for video editing (identifying time ranges for queries), while WorldSense evaluates comprehensive omnimodal understanding across 26 cognitive tasks. The benchmarks serve different purposes and are not directly comparable.

---

### 4. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis

**URL**: View paper

**Brief Assessment**

Video-MME[73] focuses on evaluating video understanding across different temporal durations (short/medium/long) with 900 videos, while WorldSense emphasizes tightly coupled audio-visual perception requiring both modalities for correct answers with 1,662 videos across 8 domains. The benchmarks differ in their core design principles and evaluation focus.

---

### 5. Foundation models for video understanding: A survey

**URL**: View paper

**Brief Assessment**

Foundation Models Survey[74] is a comprehensive survey paper that reviews existing video understanding benchmarks and models but does not present a competing benchmark. It discusses various video understanding tasks and datasets but does not claim to introduce a novel benchmark that would refute WorldSense's novelty in omnimodal evaluation.

---

### 6. Mvbench: A comprehensive multi-modal video understanding benchmark

**URL**: View paper

**Brief Assessment**

MVBench[72] focuses on temporal understanding in videos with visual-only inputs across 20 tasks, while WorldSense emphasizes omnimodal integration requiring synchronized audio-visual-text processing across 26 tasks. The benchmarks address different evaluation paradigms and modality requirements.

---

### 7. Value: A multi-task benchmark for video-and-language understanding evaluation

**URL**: View paper

**Brief Assessment**

VALUE[77] focuses on video-and-language understanding with visual and textual (subtitle) inputs across 11 datasets, but does not emphasize audio as a critical modality for omnimodal understanding. WorldSense specifically requires integrated audio-visual-text processing where audio is essential, not optional.

---

### 8. Internvideo2: Scaling foundation models for multimodal video understanding

**URL**: View paper

**Brief Assessment**

InternVideo2[70] focuses on building video foundation models for general video understanding tasks, not on creating evaluation benchmarks for omnimodal (audio-visual-text) understanding with tightly coupled modalities as WorldSense does.

---

### 9. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset

**URL**: View paper

**Brief Assessment**

VAST[78] focuses on vision-audio-subtitle-text foundation model training with automatically generated captions, not on creating a benchmark for evaluating MLLMs' real-world understanding capabilities. The candidate addresses dataset creation for pretraining, while the original addresses benchmark design for evaluation.

---

### 10. Longvideobench: A benchmark for long-context interleaved video-language understanding

**URL**: View paper

**Brief Assessment**

LongVideoBench[69] focuses on long-context video understanding with interleaved video-language inputs and referring reasoning tasks, while WorldSense emphasizes omnimodal (audio-visual-text) integration with strong audio-visual coupling. The benchmarks address different evaluation paradigms and challenges.

---

## Contribution 2: Design principles emphasizing omnimodal collaboration

**Description**: The benchmark is constructed with deliberate design principles that ensure questions require tight coupling between audio and visual modalities for correct answers. This forces models to demonstrate genuine multimodal integration rather than relying on single-modality processing, establishing a rigorous evaluation framework for omnimodal understanding.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

### 1. Onellm: One framework to align all modalities with language

**URL**: View paper

**Brief Assessment**

OneLLM[55] focuses on aligning multiple modalities to language through a unified framework and progressive training pipeline, not on designing evaluation benchmarks with tight audio-visual coupling requirements. The paper addresses model architecture and training methodology rather than evaluation task design principles.

---

### 2. Robust Audio-Visual Segmentation via Audio-Guided Visual Convergent Alignment

**URL**: View paper

**Brief Assessment**

Robust Audio Visual[21] focuses on audio-visual segmentation tasks where the goal is to segment audible objects in visual scenes. While it addresses audio-visual alignment challenges, it does not propose evaluation benchmarks or design principles for assessing omnimodal understanding in multimodal LLMs, which is the core contribution of the original paper.

---

### 3. Representation learning for semantic alignment of language, audio, and visual modalities

**URL**: View paper

**Brief Assessment**

Semantic Alignment Learning[6] focuses on contrastive learning for aligning audio, visual, and text modalities in representation space, not on designing evaluation benchmarks with questions requiring tight audio-visual coupling for correct answers.

---

### 4. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment

**URL**: View paper

**Brief Assessment**

LanguageBind[52] focuses on extending video-language pretraining to multiple modalities through language-based semantic alignment, not on designing evaluation benchmarks with tight audio-visual coupling requirements for testing models.

---

### 5. Av-superb: A multi-task evaluation benchmark for audio-visual representation models

**URL**: View paper

**Brief Assessment**

AV-SUPERB[56] focuses on evaluating audio-visual representation models across multiple tasks but does not specifically address design principles for ensuring tight coupling between modalities in question construction. The benchmark evaluates existing models rather than establishing design principles for omnimodal collaboration in evaluation tasks.

---

### 6. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners

**URL**: View paper

**Brief Assessment**

Seeing and Hearing[57] focuses on visual-audio generation using diffusion models and multimodal alignment, not on evaluation benchmark design requiring tight coupling between modalities for question-answering tasks.

---

### 7. Unsupervised audio-visual segmentation with modality alignment

**URL**: View paper

**Brief Assessment**

Unsupervised Audio Visual[51] focuses on unsupervised audio-visual segmentation for pixel-level object identification, not on evaluation benchmark design requiring tight coupling between modalities for question-answering tasks.

---

### 8. Aurelia: Test-time reasoning distillation in audio-visual llms

**URL**: View paper

**Brief Assessment**

Aurelia[54] focuses on test-time reasoning distillation for audio-visual LLMs rather than benchmark design principles. The candidate addresses reasoning enhancement through multi-agent frameworks, not evaluation task construction requiring tight audio-visual coupling.

---

### 9. DAVE: Diagnostic benchmark for Audio Visual Evaluation

**URL**: View paper

**Prior Art Analysis**

DAVE[53] demonstrates that prior work exists on designing evaluation tasks requiring tight coupling between audio and visual modalities. Both papers emphasize the necessity of integrating both modalities for correct answers, with DAVE[53] explicitly stating that 'each question requires information from both audio and visual modalities simultaneously, ensuring that neither modality alone is

sufficient.' This directly parallels the original paper's claim of 'tight coupling between audio and video, requiring models to effectively utilize the synergistic perception of omni-modality.' The design principle of forcing genuine multimodal integration rather than single-modality processing was already established in DAVE[53].

**Evidence**

Evidence 1 - **Rationale**: Both papers explicitly describe the same design principle: requiring tight coupling between audio and visual modalities where neither modality alone is sufficient for correct answers. - **Original**: collaboration of omni-modality, we design the evaluation tasks to feature a strong coupling of audio and video, requiring models to effectively utilize the synergistic perception of omni-modality - **Candidate**: dave is designed around a key principle: each question requires information from both audio and visual modalities simultaneously, ensuring that neither modality alone is sufficient

Evidence 2 - **Rationale**: Both papers describe the same evaluation framework where removing either modality results in failure, establishing rigorous assessment of multimodal integration. - **Original**: the benchmark emphasizes the joint processing of audio and visual modalities, as illustrated in figure 1. each question requires both modalities for accurate response-removing either results in failure-enabling rigorous assessment of a model's capacity for integrated sensory understanding - **Candidate**: unlike existing benchmarks that suffer from modality bias, dave explicitly requires information from both auditory and visual modalities, ensuring that neither modality alone is sufficient for correctly answering questions

---

### 10. Study of subjective and objective quality assessment of audio-visual signals

**URL**: View paper

**Brief Assessment**

Subjective Objective Quality[58] focuses on quality assessment of audio-visual signals through distortion analysis and quality prediction models, not on designing evaluation tasks that require tight coupling between modalities for comprehension and reasoning.

---

## Contribution 3: Comprehensive evaluation revealing limitations of current MLLMs

**Description**: The authors conduct extensive experiments on state-of-the-art models, revealing that even the best proprietary model achieves only 65.1% accuracy while open-source models perform near chance level. Through ablation studies and failure analysis, they identify key factors influencing performance and provide actionable insights for improving omnimodal understanding in future models.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

### 1. Generative multimodal models are in-context learners

**URL**: View paper

**Brief Assessment**

Generative In Context[60] focuses on evaluating in-context learning abilities of generative multimodal models across understanding and generation tasks, not specifically on real-world omnimodal understanding with audio-visual integration that WorldSense emphasizes.

---

### 2. Lmms-eval: Reality check on the evaluation of large multimodal models

**URL**: View paper

**Brief Assessment**

LMMS-Eval[68] focuses on standardizing evaluation pipelines and addressing contamination issues across diverse benchmarks, rather than conducting comprehensive experiments on omnimodal understanding with audio-visual integration as in the original paper.

---

### 3. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks

**URL**: View paper

**Brief Assessment**

VisualWebArena[59] focuses on evaluating multimodal agents on web navigation tasks with visual grounding, not on general omnimodal understanding across diverse real-world scenarios with audio-visual integration as in the original paper.

---

### 4. Probing multimodal llms as world models for driving

**URL**: View paper

**Brief Assessment**

Probing World Models[65] focuses specifically on evaluating MLLMs for autonomous driving scenarios (ego-motion, traffic, trajectory planning), while the original paper evaluates omnimodal understanding across diverse real-world domains (8 categories, 67 subcategories). The driving-specific focus and different evaluation methodology do not challenge the novelty of a comprehensive benchmark for general real-world omnimodal understanding.

---

### 5. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans?

**URL**: View paper

**Brief Assessment**

MME-RealWorld[63] focuses on high-resolution image understanding across specific real-world domains (OCR, remote sensing, autonomous driving, etc.), while the original paper evaluates omnimodal video understanding requiring audio-visual integration. The candidate's evaluation is image-based and does not address the audio-visual coupling challenges central to the original work.

---

### 6. Tracking meets large multimodal models for driving scenario understanding

**URL**: View paper

**Brief Assessment**

Tracking Multimodal Driving[61] focuses on integrating tracking information into multimodal models for autonomous driving scenarios, not on comprehensive evaluation of MLLMs across diverse real-world domains. The candidate addresses a different technical problem (tracking-enhanced driving understanding) rather than evaluating model limitations across broad omnimodal scenarios.

---

### 7. Touchstone: Evaluating vision-language models by language models

**URL**: View paper

**Brief Assessment**

Touchstone[66] focuses on evaluating vision-language models through conversational quality assessment using text-based LLM judges, not on omnimodal (audio-visual-text) understanding in real-world scenarios like the original paper.

---

### 8. Videoautoarena: An automated arena for evaluating large multimodal models in video analysis through user simulation

**URL**: View paper

**Brief Assessment**

VideoAutoArena[67] focuses on automated arena-style evaluation through user simulation and peer battles for video analysis, while the original paper conducts comprehensive experiments on omnimodal understanding (audio-visual-text integration). The evaluation methodologies and focus areas differ substantially.

---

### 9. Llava-critic: Learning to evaluate multimodal models

**URL**: View paper

**Brief Assessment**

Llava-Critic[62] focuses on training models to evaluate multimodal outputs (LMM-as-a-judge), not on conducting comprehensive evaluations of existing models' real-world understanding capabilities across diverse scenarios.

---

### 10. Hallucination of Multimodal Large Language Models: A Survey

**URL**: View paper

**Brief Assessment**

Hallucination Survey[64] focuses on hallucination phenomena in MLLMs (inconsistencies between outputs and visual content), while the original paper evaluates omnimodal understanding capabilities across audio-visual-text modalities in real-world scenarios. These are distinct evaluation objectives with different scopes.

---

## Appendix: Text Similarity Detection

Textual similarity detection checked 33 papers and found 3 similarity segment(s) across 2 paper(s).

The following **2 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

### 1. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans?

**Detected in**: Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

### 2. Daily-Omni: Towards Audio-Visual Reasoning with Temporal Alignment across Modalities

**Detected in**: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

## References


- [0] WorldSense: Evaluating Real-world Omnimodal Understanding for Multimodal LLMs View paper
- [1] Unified video-language pre-training with synchronized audio View paper
- [2] Audio-centric video understanding benchmark without text shortcut View paper
- [3] Multimodal alignment and fusion: A survey View paper
- [4] Cross-modal prompts: Adapting large pre-trained models for audio-visual downstream tasks View paper
- [5] OmniEval: A Benchmark for Evaluating Omni-modal Models with Visual, Auditory, and Textual Inputs View paper
- [6] Representation learning for semantic alignment of language, audio, and visual modalities View paper
- [7] CineBrain: A Large-Scale Multi-Modal Brain Dataset During Naturalistic Audiovisual Narrative Processing View paper
- [8] Audio-visual scene analysis with self-supervised multisensory features View paper
- [9] Watch and Listen: Understanding Audio-Visual-Speech Moments with Multimodal LLM View paper
- [10] Intra-modal and Cross-modal Synchronization for Audio-visual Deepfake Detection and Temporal Localization View paper
- [11] X-Streamer: Unified Human World Modeling with Audiovisual Interaction View paper
- [12] Daily-Omni: Towards Audio-Visual Reasoning with Temporal Alignment across Modalities View paper
- [13] Audio-Sync Video Generation with Multi-Stream Temporal Control View paper
- [14] Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization View paper
- [15] Ma-avt: Modality alignment for parameter-efficient audio-visual transformers View paper
- [16] Av-link: Temporally-aligned diffusion features for cross-modal audio-video generation View paper
- [17] OmniMMI: A Comprehensive Multi-modal Interaction Benchmark in Streaming Video Contexts View paper
- [18] Long-Video Audio Synthesis with Multi-Agent Collaboration View paper
- [19] Quality over quantity? llm-based curation for a data-efficient audio-video foundation model View paper
- [20] Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions View paper
- [21] Robust Audio-Visual Segmentation via Audio-Guided Visual Convergent Alignment View paper
- [22] SVLTA: Benchmarking Vision-Language Temporal Alignment via Synthetic Video Situation View paper
- [23] Diverse and aligned audio-to-video generation via text-to-video model adaptation View paper
- [24] Controllable Video-to-Music Generation with Multiple Time-Varying Conditions View paper
- [25] Omni-R1: Reinforcement Learning for Omnimodal Reasoning via Two-System Collaboration View paper
- [26] Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog View paper
- [27] Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue View paper
- [28] Vggsounder: Audio-visual evaluations for foundation models View paper
- [29] Unified multisensory perception: Weakly-supervised audio-visual video parsing View paper
- [30] 2DP-2MRC: 2-Dimensional Pointer-based Machine Reading Comprehension Method for Multimodal Moment Retrieval View paper
- [31] AlignVSR: Audio-visual cross-modal alignment for visual speech recognition View paper
- [32] JavisGPT: A Unified Multi-modal LLM for Sounding-Video Comprehension and Generation View paper
- [33] Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos View paper
- [34] Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis View paper

- [35] Coavt: A cognition-inspired unified audio-visual-text pre-training model for multimodal processing View paper
- [36] Internlm-xcomposer2. 5-omnilive: A comprehensive multimodal system for long-term streaming video and audio interactions View paper
- [37] AVoCaDO: An Audiovisual Video Captioner Driven by Temporal Orchestration View paper
- [38] CLIP-Powered TASS: Target-Aware Single-Stream Network for Audio-Visual Question Answering View paper
- [39] Answering Diverse Questions via Text Attached with Key Audio-Visual Clues View paper
- [40] Aligning What Matters: Masked Latent Adaptation for Text-to-Audio-Video Generation View paper
- [41] A Bridge from Audio to Video: Phoneme-Viseme Alignment Allows Every Face to Speak Multiple Languages View paper
- [42] Harmonizing Sight and Sound: The Impact of Auditory Emotional Arousal, Visual Variation, and Their Congruence on Consumer Engagement in Short Video Marketing View paper
- [43] Enhancing Cross-Modal Understanding for Audio Visual Scene-Aware Dialog Through Contrastive Learning View paper
- [44] LongCat-Flash-Omni Technical Report View paper
- [45] Video-to-audio generation with hidden alignment View paper
- [46] Interactive Retrieval System for Multi-Stream Collections: multiXview at CASTLE 2025 Interactive Grand Challenge View paper
- [47] AVS-Mamba: Exploring Temporal and Multi-Modal Mamba for Audio-Visual Segmentation View paper
- [48] AS-Net: active speaker detection using deep audio-visual attention View paper
- [49] Multi-dimensional fusion: transformer and GANs-based multimodal audiovisual perception robot for musical performance art View paper
- [50] Learning to unify audio, visual and text for audio-enhanced multilingual visual answer localization View paper
- [51] Unsupervised audio-visual segmentation with modality alignment View paper
- [52] Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment View paper
- [53] DAVE: Diagnostic benchmark for Audio Visual Evaluation View paper
- [54] Aurelia: Test-time reasoning distillation in audio-visual llms View paper
- [55] Onellm: One framework to align all modalities with language View paper
- [56] Av-superb: A multi-task evaluation benchmark for audio-visual representation models View paper
- [57] Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners View paper
- [58] Study of subjective and objective quality assessment of audio-visual signals View paper
- [59] Visualwebarena: Evaluating multimodal agents on realistic visual web tasks View paper
- [60] Generative multimodal models are in-context learners View paper
- [61] Tracking meets large multimodal models for driving scenario understanding View paper
- [62] Llava-critic: Learning to evaluate multimodal models View paper
- [63] Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? View paper
- [64] Hallucination of Multimodal Large Language Models: A Survey View paper
- [65] Probing multimodal llms as world models for driving View paper
- [66] Touchstone: Evaluating vision-language models by language models View paper
- [67] Videoautoarena: An automated arena for evaluating large multimodal models in video analysis through user simulation View paper
- [68] Lmms-eval: Reality check on the evaluation of large multimodal models View paper
- [69] Longvideobench: A benchmark for long-context interleaved video-language understanding View paper
- [70] Internvideo2: Scaling foundation models for multimodal video understanding View paper
- [71] Vidi: Large multimodal models for video understanding and editing View paper
- [72] Mvbench: A comprehensive multi-modal video understanding benchmark View paper
- [73] Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis View paper
- [74] Foundation models for video understanding: A survey View paper
- [75] Internvid: A large-scale video-text dataset for multimodal understanding and generation View paper
- [76] Iv-bench: A benchmark for image-grounded video perception and reasoning in multimodal llms View paper
- [77] Value: A multi-task benchmark for video-and-language understanding evaluation View paper
- [78] Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset View paper