AVID: BENCHMARKING AUDIO-VISUAL INCONSISTENCY UNDERSTANDING FOR OMNI-MODAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal large language models demonstrate strong content understanding in tasks such as captioning and question answering, yet their ability to understand fine-grained audio-visual inconsistencies remains largely unevaluated. To bridge this gap, we present AVID, a benchmark for audio-visual inconsistency understanding comprising 11.2K long-form videos with 39.4K annotated inconsistency events and 78.7K derived segment clips, with source-video-disjoint training and test splits. It supports detection, temporal grounding, classification, and reasoning across eight fine-grained inconsistency categories, including temporal shifts, semantic contradictions, and environmental conflicts. Compared with prior benchmarks focused on aligned events, forgery detection, or controlled multiple-choice diagnostics, it evaluates the localization and explanation of multiple inconsistencies in long videos, spanning low-level physical mismatches and high-level semantic and contextual conflicts. The benchmark is constructed through a scalable pipeline that combines content-type-aware temporal segmentation with an agentic framework for category selection and diverse audio-visual conflict injection. Modality ablations and processed-consistent controls show that strong performance requires cross-modal comparison and cannot be explained by manipulation artifacts. Extensive experiments and analyses of open- and closed-source models reveal persistent limitations in fine-grained inconsistency understanding. Using the training split, we develop AVID-Qwen, which improves over its base model across detection, temporal grounding, classification, and reasoning. Adding examples from the training split to general training data improves inconsistency detection while retaining general omni-modal capabilities.

1 INTRODUCTION

Omni-modal large language models (Gemini Team, Google DeepMind, 2025; Xiaomi AI Lab, 2025; Xu et al., 2025) achieve strong performance on video captioning, visual question answering (VQA), and temporal action localization by jointly modeling visual, auditory, and textual signals (Geng et al., 2025; Goel et al., 2024; Zhang et al., 2023; Fu et al., 2025). However, one fundamental cognitive ability remains largely overlooked in current evaluation: the understanding of audio-visual *inconsistency* (Ye et al., 2026). For instance, humans can recognize mismatches between a speaker's apparent identity and vocal characteristics, or contradictions between spoken descriptions and visual events (Figure 1). Such judgments require comparing evidence across modalities and interpreting the source of disagreement, yet the extent to which models possess this capability remains unclear.

Despite its importance for trustworthy multimodal understanding, hallucination detection, and audio-visual reasoning (Huang et al., 2024; Wang et al., 2023), this capability is not comprehensively evaluated by existing benchmarks. Mainstream grounding benchmarks (Tian et al., 2018; Geng et al., 2023; 2025) evaluate the localization of audio-visual events under an assumption of cross-modal consistency, while deepfake datasets (Yang et al., 2023; Cai et al., 2022; 2025) focus on detecting manipulated content and localizing forgeries. DAVE (Radevski et al., 2025) uses controlled multiple-choice questions to evaluate sound–action temporal association, sound absence detection, and sound discrimination. This motivates a benchmark that comprehensively evaluates fine-grained audio-visual inconsistency understanding across physical, semantic, and contextual

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

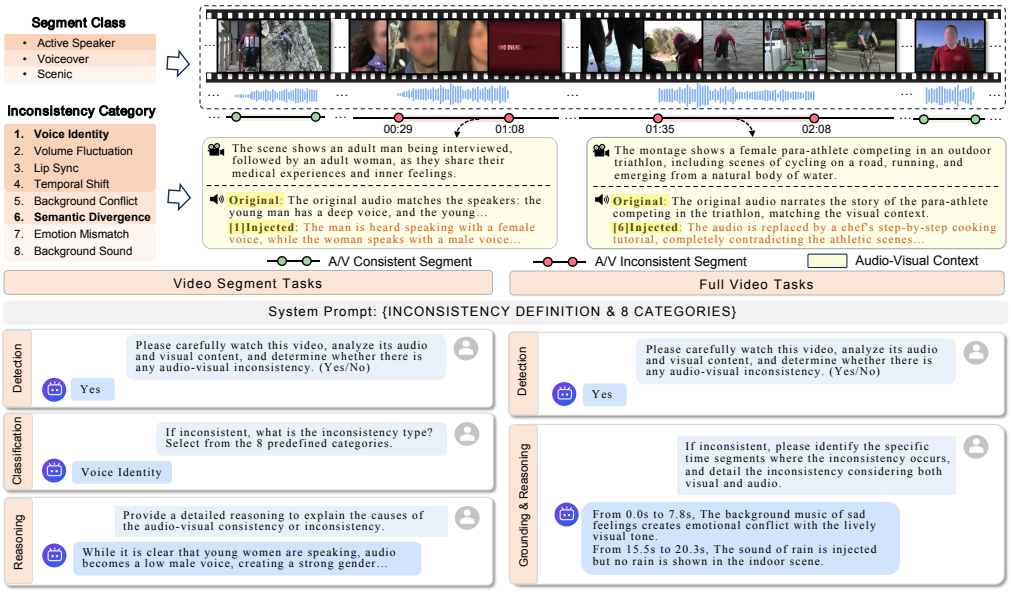


Figure 1: Overview of AVID, a benchmark for audio-visual inconsistency understanding.

conflicts, covering detection, temporal localization, categorization, and explanation at both segment and full-video scales (Table 1).

To address these limitations, we introduce **AVID**, a systematic benchmark for long-form audio-visual inconsistency understanding. We develop a scalable framework of three stages: class-aware temporal segmentation into Active Speaker, Voiceover, and Scenic categories; strategy-guided inconsistency planning via an LLM agent; and controlled conflict injection through five specialized injectors, with human oversight (§3). AVID comprises 11.2K full videos (avg. 235.5 s) with 39.4K annotated inconsistency events and 78.7K segment clips (avg. 14.8 s), with training and test splits disjoint at the source-video level. The benchmark supports evaluation across detection, temporal grounding, classification, and reasoning over 8 fine-grained inconsistency categories (Table 1). We first examine whether benchmark performance reflects audio-visual inconsistency understanding through modality ablations and processed-consistent controls. We then evaluate representative open-source and closed-source omni-modal models on the test split. Both groups exhibit limitations: without AVID-specific fine-tuning, the highest segment detection accuracy reaches 69.7%, and the highest full-video mIoU reaches 26.2%. Fine-grained analyses examine category-level confusion, the balance between recall and false positives, and the trade-off between grounding coverage and precision, revealing distinct error patterns across models. We also use the AVID training split to develop **AVID-Qwen** through two-stage fine-tuning of Qwen3-Omni, improving detection, temporal grounding, classification, and reasoning over the base model. In a separate experiment, adding 20K segment examples from this split to a general omni-modal training mixture raises inconsistency detection accuracy from 51.5% to 57.0% while preserving general omni-modal performance relative to the base model, further demonstrating the utility of AVID as a complementary sub-task dataset. Our contributions are as follows:

- We design a scalable framework for constructing audio-visual inconsistencies in videos, comprising content-aware segmentation, strategy-guided inconsistency planning, and controlled conflict injection with human oversight.
- We introduce AVID, a benchmark comprising 11.2K full videos and 78.7K derived clips with training and test splits, supporting inconsistency detection, eight-category classification, temporal grounding, and free-form reasoning. Together, these tasks provide a comprehensive assessment of fine-grained audio-visual inconsistency understanding.
- We systematically evaluate existing open-source and closed-source omni-modal models and conduct fine-grained analyses, revealing their limitations in audio-visual inconsistency understanding. We also provide an effective training split and develop AVID-Qwen through two-stage fine-tuning, improving inconsistency understanding over its base model.

Dataset	Anno.	Video Form	#Videos	Avg. Len	Caption	A/V Inc.	#Inc. Cat.	Inc. Reasoning	Temp. Loc.
Grounding / Localization Benchmarks									
InternVid (Wang et al., 2024)	G	Segment	234M	11.7 s	V	✗	–	–	✗
Panda-70M (Chen et al., 2024)	G	Segment	70.8M	8.5 s	V	✗	–	–	✗
AudioCaps (Kim et al., 2019)	M	Segment	39.6K	10 s	A	✗	–	–	✗
WavCaps (Mei et al., 2024)	G	Segment	403K	67.6 s	A	✗	–	–	✗
ACAV100M (Lee et al., 2021)	G	Segment	100M	10 s	–	✗	–	–	✗
VALOR (Liu et al., 2024a)	M	Segment	1.18M	10 s	VA	✗	–	–	✗
VAST (Chen et al., 2023)	G	Segment	27M	–	VAS	✗	–	–	✗
AVE (Tian et al., 2018)	M	Segment	4,143	10 s	VA	✗	–	–	✓
UnAV-100 (Geng et al., 2023)	M	Full Video	10,790	42.1 s	–	✗	–	–	✓
ActivityNet Caps (Krishna et al., 2017)	M	Full Video	20K	180 s	V	✗	–	–	✓
Charades-STA (Gao et al., 2017)	G+M	Full Video	10K	30 s	V	✗	–	–	✓
LongVALE (Geng et al., 2025)	G+M	Full Video	8,411	235 s	VAS	✗	–	–	✓
Deepfake / Synchronization Benchmarks									
FakeAVCeleb (Khalid et al., 2021)	G+M	Segment	20,000	7.8 s	–	✓	3	✗	✗
LAV-DF (Cai et al., 2022)	G	Segment	136K	–	–	✓	3	✗	✓
LipFD / AVLips (Liu et al., 2024b)	G	Segment	340K	–	–	✓	1	✗	✗
AV-Deepfake1M++ (Cai et al., 2025)	G	Segment	2M	8.2 s	–	✓	3	✗	✓
Audio-Visual Question Answering / Diagnostics									
DAVE (Radevski et al., 2025)	G+M	Segment	2,426	–	VA	–	–	✗	✗
Audio-Visual Inconsistency Understanding									
AVID (Ours)	G+M	Segment Full Video	**78,722** **11,200**	**14.8 s** **235.5 s**	VAS	✓	**8**	✓	✓

Table 1: Comparison with representative audio-visual benchmarks. Anno.: annotation method (G: generated, M: manual). A/V Inc.: whether audio-visual inconsistency is an explicit evaluation target. #Inc. Cat.: number of inconsistency categories. Inc. Reasoning: free-form inconsistency explanations. Temp. Loc.: temporal localization evaluated as an output task. V: visual, A: audio, S: speech.

2 Related Work

Audio-Visual Benchmarking. AVE (Tian et al., 2018) localizes events that are both audible and visible, while UnAV-100 (Geng et al., 2023) extends this task to multiple event instances in untrimmed videos. LongVALE (Geng et al., 2025) further supports audio-visual-language event localization and description in long videos. These benchmarks evaluate the recognition, localization, or description of corresponding audio-visual events; their categories describe event content rather than types of cross-modal conflict. DAVE (Radevski et al., 2025) uses controlled multiple-choice questions to evaluate sound–action temporal association, sound absence detection, and sound discrimination. Its overlaid sounds need not conflict with the visual scene, and its three question types do not define inconsistency categories. AVID jointly evaluates the detection, localization, classification, and explanation of audio-visual inconsistencies at segment and full-video scales (Table 1).

Deepfake and Inconsistency Detection. FakeAVCeleb (Khalid et al., 2021) benchmarks multimodal deepfake detection, and LAV-DF (Cai et al., 2022) provides temporally localized audio-visual manipulations for forgery localization. Methods such as AVFF (Oorloff et al., 2024) and AVoiD-DF (Yang et al., 2023) exploit cross-modal correspondences or inconsistencies to detect forgeries. These works target authenticity and forgery localization. AVID evaluates inconsistency understanding across eight fine-grained physical, semantic, and contextual categories, with modality ablations and processed-consistent controls assessing the influence of synthesis artifacts.

3 The AVID Benchmark

To construct AVID, we propose a scalable pipeline consisting of three stages: (i) data sourcing and class-aware segmentation; (ii) strategy planning; and (iii) inconsistency injection. The overall workflow is illustrated in Figure 2, with implementation details provided in Appendix A.

3.1 Data Sourcing and Class-Aware Segmentation

We build AVID on top of videos from the LongVALE collection (Geng et al., 2025), which are curated for verified semantic consistency between audio and visual content. Starting from verified-consistent source videos is critical: it ensures that any inconsistency in the resulting benchmark originates exclusively from controlled injection, enabling unambiguous ground-truth annotation without confounding pre-existing mismatches. Since inconsistency patterns are inherently class-dependent,

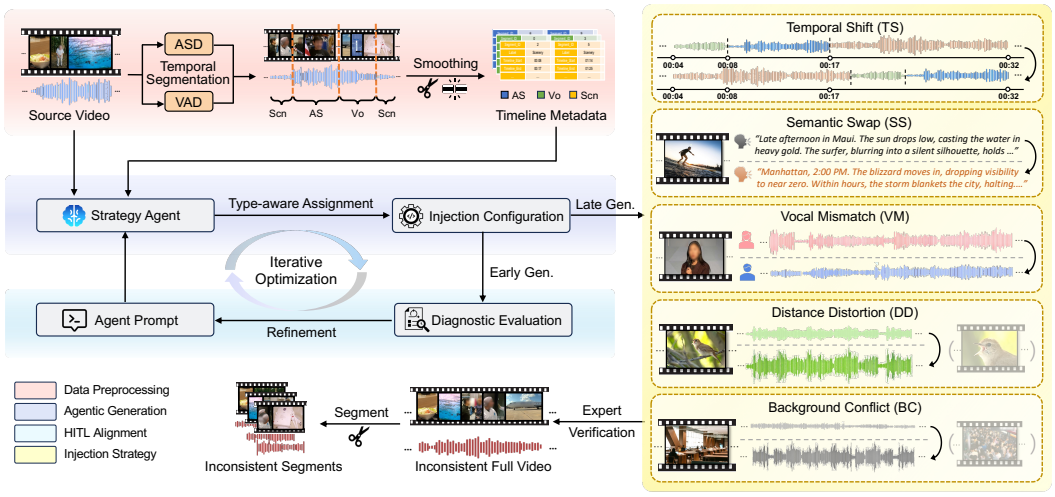


Figure 2: Overview of the AVID construction pipeline, consisting of three stages: (1) data preprocessing with temporal segmentation and class-aware labeling, (2) agentic generation with strategy planning and five specialized inconsistency injectors, and (3) human-in-the-loop alignment.

we first segment each video into class-aware units prior to any manipulation. Using a four-step pipeline—speech activity detection (Silero VAD (Silero Team, 2021)), active-speaker proxy detection (OpenCV Haar Cascade), rule-based class assignment, and temporal fusion—we classify each segment into one of three semantic classes based on two attributes: whether speech is present and whether a visible speaker appears (Appendix Table 6). **Class 1 (Active Speaker)**: human speech is present and the speaker is visible on screen, enabling evaluation of lip-sync, voice identity, and temporal alignment. **Class 2 (Voiceover)**: human speech is present but the speaker is not visible, suitable for evaluating semantic alignment between narration and visual content. **Class 3 (Scenic)**: no human speech is present, suitable for evaluating environmental and affective audio-visual consistency. Based on these classes, we define a taxonomy of 8 inconsistency categories (Appendix Table 8). This class-aware design ensures that injected inconsistencies are semantically grounded in the segment context, enabling evaluation of genuine cross-modal reasoning.

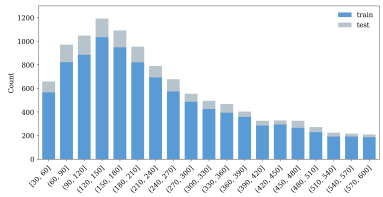
3.2 Strategy Planning

After segmentation, we introduce a **Strategy Planning Agent** (Gemini 3.1 Pro) to determine *whether* each segment is suitable for injection and *which* inconsistency type to introduce, rather than applying injectors indiscriminately (Figure 2, Stage 2).

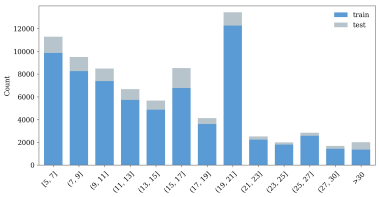
The agent takes as input the temporal segments produced by Stage 1 along with their class labels, and reads the corresponding video to understand the audio-visual content. Based on this shared context, the agent makes two progressive decisions. First, **feasibility screening**: the agent assesses whether each segment contains sufficient and unambiguous audio-visual content to support a perceptible inconsistency. Second, **category selection**: for each feasible segment, the agent selects the most contextually appropriate inconsistency type from a class-conditional candidate set (4 types for Class 1, 2 for Class 2, 2 for Class 3; full mapping in Appendix), while balancing category diversity across the video. For each selected segment, the agent produces an injection plan that specifies the temporal boundaries, inconsistency category, injector parameters, and a natural-language reasoning describing the expected audio-visual conflict. The strategy prompt is iteratively refined through human feedback, and all generated strategies undergo manual review before execution.

3.3 Inconsistency Injection

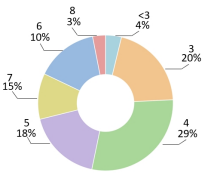
Given the injection plan produced by the Strategy Planning Agent, an Execution Agent parses each entry and routes the corresponding segment to one of five specialized injectors according to the specified inconsistency category (Figure 2, Stage 2): (1) **Temporal**—applies audio time-axis off-



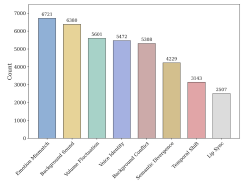
(a) Video duration distribution (train/test)



(b) Segment duration distribution (train/test)



(c) Inconsistency events per video



(d) Inconsistency categories



(e) Examples of 8 audio-visual inconsistency categories.

Figure 3: Comprehensive dataset statistics: (a) video duration distribution for training and test sets, (b) segment duration distribution, (c) pie chart of inconsistency events per video, (d) counts of the 8 audio-visual inconsistency categories, and (e) examples of 8 audio-visual inconsistency categories.

set; (2) **Semantic**—replaces speech with contradictory TTS content; (3) **Identity**—modifies vocal characteristics (pitch, formant, timbre); (4) **Spatial**—simulates distance-based volume attenuation; and (5) **Background**—replaces environmental audio with scene-contradictory ambience or affective music. The class-conditional mapping between segment classes, inconsistency categories, and injectors is detailed in Appendix Table 8.

Crucially, all injectors operate exclusively on the audio stream; the original video frames, resolution, and duration are preserved throughout. This design makes the target label a relation between the unchanged visual evidence and the modified audio, while avoiding visual manipulation artifacts. Modality ablation and processed-consistent controls empirically test whether audio-side traces nevertheless provide a shortcut. Implementation details for each injector are provided in Appendix A.3.

3.4 Human Effort and Subset Split

Human oversight is incorporated at two stages. **Stage I: Strategy-level filtering.** Annotators iteratively review sampled strategy outputs, flag unreasonable plans, and provide feedback to refine the strategy prompt (final prompt in Appendix Figure 7). After every 50 generated strategies, experts review them until fewer than 3 are flagged per batch (95% acceptance rate). Only approved strategies are dispatched to injectors. **Stage II: Post-construction quality control.** We manually inspect the constructed videos against the principle of human verifiability: given the 8 inconsistency category definitions, a human viewer should be able to determine consistency and identify the category. Videos with poor quality or ambiguous inconsistencies are filtered out. Inter-annotator agreement studies confirm substantial agreement (Cohen's $\kappa > 0.75$). Through this pipeline, we construct 8,400 full videos with inconsistency events (39,361 inconsistent segments). The full-video set contains 11,200 videos (3:1 inconsistent-to-consistent ratio), split into 9,639/1,561 for training and testing. The segment set contains 78,722 clips (1:1 ratio), split into 68,088/10,634.

3.5 Statistical Analysis

Figure 3 presents comprehensive statistics. The segment-level set (78.7K clips, 5–20 s) supports fine-grained inconsistency detection, classification, and reasoning over self-contained evidence. The full-video set (11.2K videos, 30–600 s) requires identifying multiple inconsistency events with temporal grounding and dense reasoning—a substantially more challenging setting. Inconsistency events are densely distributed: 78% of videos contain 3–6 events, and 18% contain more than 7. The

8 inconsistency categories are balanced across both sets (Figure 3d). The full-video annotations naturally align with temporal video grounding (TVG) and dense video captioning (DVC) formulations. We also release standardized train/test splits in both formats.

4 AVID-QWEN: A STRONG OMNI-MODAL BASELINE

To demonstrate the learnability of AVID and provide a reproducible reference point, we build a strong baseline by fine-tuning an open-weight omni-modal model on our benchmark data (Figure 4). We denote this model AVID-Qwen.

4.1 BACKBONE MODEL

We adopt **Qwen3-Omni-30B-A3B-Instruct** as the backbone, an omni-modal LLM that natively processes interleaved video frames and audio to generate textual responses. It consists of a vision encoder, an audio encoder, and an LLM backbone that reasons over concatenated visual and audio tokens. Architecture details and preprocessing configurations are provided in Appendix B. We design task-specific prompts for two evaluation scenarios: (i) segment-level detection, classification, and reasoning; and (ii) full-video detection with temporal grounding and dense reasoning.

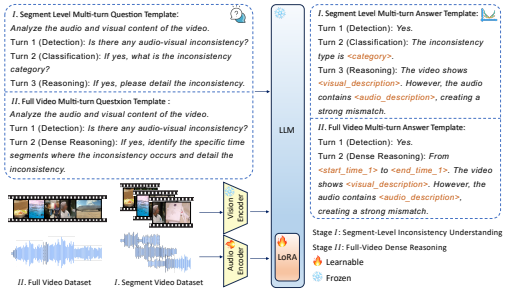


Figure 4: The AVID-Qwen architecture for omni-modal audio-visual inconsistency detection and reasoning.

4.2 TWO-STAGE FINE-TUNING

Zero-shot omni-models struggle with fine-grained temporal grounding in long-form videos, while training on full-video data alone yields suboptimal localization (Table 5). We therefore adopt a **progressive two-stage strategy** (Figure 4) that first builds local detection capability and then extends it to multi-event grounding. In **Stage 1**, the model is fine-tuned on segment-level data (5–20 s clips) to detect inconsistencies, classify their categories, and generate reasoning. In **Stage 2**, fine-tuning continues on full-video data (30–600 s), where the model learns to localize multiple inconsistency events with temporal boundaries and produce dense reasoning. The resulting AVID-Qwen serves as a transparent reference for comparing omni-models under a unified evaluation protocol.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Compared Models. We evaluate both closed-source and open-source omni-models, categorized as follows: Closed-source models: Gemini 2.5 Pro, Gemini 3.1 Pro, and MiMo-V2-Omni (accessed via API). Open-source models: Qwen3-Omni, OLA (Liu et al., 2025), Video-SALMONN 2 (Sun et al., 2024) (7B, 72B), Gemma4 E4B (Google DeepMind, 2026), and ARC-Hunyuan-Video (Ge et al., 2025) (self-hosted). Our model: AVID-Qwen, built on Qwen3-Omni and fine-tuned on both segment-level and full-video data (Stage 1 + Stage 2) as described in Section 4.

Evaluation Protocol. We evaluate on two complementary tasks: (i) **segment-level** (detection, classification, reasoning) and (ii) **full-video** (detection, temporal grounding, reasoning). The evaluation follows a coarse-to-fine pipeline: detect first, then classify or ground, and finally reason. Classification, grounding, and reasoning are evaluated only on true positives.

Metrics. We report Accuracy for detection and classification, BLEU-4 (Papineni et al., 2002) / ROUGE-L (Lin, 2004) / METEOR (Banerjee & Lavie, 2005) for reasoning, Recall@1 at IoU $\in \{0.3, 0.5, 0.7\}$ and mIoU for temporal grounding, and SODA-m (Fujita et al., 2020) for dense understanding. Classification, grounding, and reasoning metrics are conditioned on correctly detected samples (TP); recall- and precision-denominator variants are additionally reported. Model configurations, prompt templates, metric definitions, and metric variants are provided in Appendix C.

Model	Segment					Full-Video								
	Detect	Cla.	Reasoning			Detect	Grounding				Reasoning			Dense
	Acc	Acc	B	R	M	Acc	R@0.3	R@0.5	R@0.7	mIoU	B	R	M	S_M
Arc-Hunyuan-Video	41.6	11.1	1.4	16.3	17.0	52.3	9.8	6.2	2.8	6.4	0.01	1.55	0.73	0.57
OLA	48.1	7.9	1.3	15.4	11.9	59.3	14.3	7.1	2.8	8.8	0.17	3.64	1.94	1.20
SALMONN 2 -7B	49.9	50.0	1.5	20.9	20.8	25.1	25.0	25.0	0.0	13.8	0.16	6.45	4.95	0.38
SALMONN 2 -72B	51.5	32.3	1.62	16.5	17.1	25.0	-	-	-	-	-	-	-	-
Gemma4 E4B	49.1	53.8	1.8	22.1	19.6	-	-	-	-	-	-	-	-	-
Qwen3-Omni	52.6	55.0	2.2	21.7	23.8	75.3	12.3	8.3	4.9	9.1	0.25	3.90	2.48	1.48
MiMo-V2-Omni	60.5	53.6	2.6	22.6	22.8	63.9	23.3	22.2	18.9	19.5	0.53	5.50	5.08	5.37
Gemini 2.5 Pro	68.2	56.4	2.2	20.6	22.1	**86.9**	28.0	25.9	23.8	24.3	0.64	6.46	5.49	6.15
Gemini 3.1 Pro	**69.7**	**57.1**	2.5	21.2	22.6	81.6	30.9	28.1	**25.4**	26.2	0.58	7.04	5.12	5.83
AVID-Qwen	61.3	55.5	**6.2**	**26.5**	**25.9**	78.2	**55.2**	**39.2**	19.1	**36.1**	**2.73**	**15.2**	**15.0**	**7.47**

Table 2: Evaluation of omni-modal models on AVID. **Cla.**: classification. **R@**α: Recall@1 at IoU α. **B/R/M**: BLEU-4, ROUGE-L, METEOR. S_M: SODA-m. All values are percentages ($\times 100$).

5.2 Main Results

Table 2 summarizes the results on both segment-level and full-video tasks.

Overall Trends. Closed-source models generally achieve stronger detection performance than open-source models. The highest detection accuracies among closed-source models are 69.7% at the segment level and 86.9% at the full-video level, compared with 52.6% and 75.3%, respectively, among open-source models without AVID-specific fine-tuning. Open-source models also vary substantially across tasks. Qwen3-Omni performs relatively well in both evaluation settings, whereas OLA and ARC-Hunyuan-Video obtain low classification and reasoning scores. Video-SALMONN 2 rarely predicts inconsistencies, making its classification accuracy sensitive to the small set of true positives on which it is computed (Appendix C.4). AVID-Qwen achieves segment-level and full-video BLEU-4 scores that are 2.8 and 10.9 times those of its Qwen3-Omni base model. It also reaches 36.1% mIoU and 7.47 SODA-m, compared with 26.2% and 5.83 for Gemini 3.1 Pro. This comparison involves a model fine-tuned on AVID and a model evaluated without task-specific fine-tuning. The improvements over the base model demonstrate the utility of AVID supervision for learning fine-grained audio-visual inconsistency understanding.

Segment-level Performance. The best closed-source model achieves 69.7% detection accuracy, while open-source models obtain approximately 41% to 53%, indicating a substantial gap in inconsistency detection. Among correctly detected samples, Qwen3-Omni achieves 55.0% classification accuracy, approaching the 57.1% obtained by Gemini 3.1 Pro. In comparison, OLA and ARC-Hunyuan-Video achieve only 7.9% and 11.1%, respectively, indicating difficulty distinguishing inconsistency categories even when detection succeeds. For reasoning, Qwen3-Omni obtains a METEOR score of 23.8%, comparable to or higher than those of the evaluated closed-source models, while OLA and ARC-Hunyuan-Video score below 18%. These results reveal uneven capabilities across detection, fine-grained classification, and explanation generation.

Full-video Performance. Gemini 2.5 Pro achieves 86.9% detection accuracy on full videos. However, localizing and explaining the detected inconsistencies remain challenging. The highest mIoU among models without AVID-specific fine-tuning is 26.2%, and most evaluated open-source models obtain less than 10%. All models evaluated without task-specific fine-tuning score below 0.7 on full-video BLEU-4. Thus, relatively high video-level detection accuracy does not imply precise temporal localization or accurate multi-event explanations. Gemma4 E4B is excluded from this setting because of its 30-second audio input limit. Additional analyses using sentence-level semantic similarity and category-wise results are provided in Appendix C.6.

5.3 Benchmark Validity and Training Utility

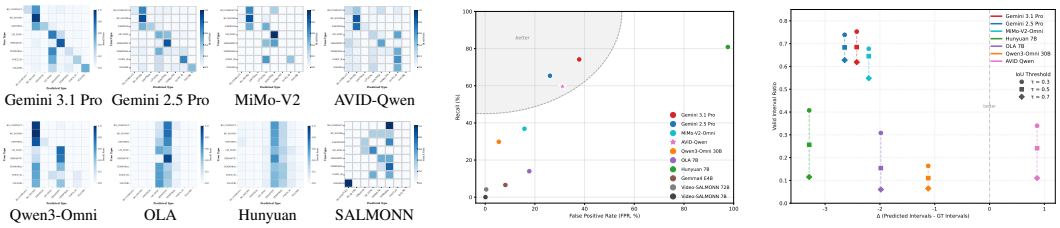
Does AVID evaluate audio-visual inconsistency understanding? Because our injectors modify audio while preserving video, models may detect synthesis artifacts without understanding audio-visual inconsistencies. We examine this possibility with two complementary controls (Table 3).

Model	Detection Acc. A+V	A	V	Artifact FPR Clean	TTS	Δ
Gemini 3.1 Pro	69.7	55.8	51.2	36.0	36.5	+0.5
AVID-Qwen	61.3	53.3	50.6	31.5	34.0	+2.5

Table 3: Detection accuracy by input modality and false-positive rates on original (Clean) and semantically consistent TTS-replaced clips. Values are percentages; Δ: FPR change in percentage points.

SFT Data	DO	WS	AVID
Base	62.32	46.00	51.5
+ General	**63.83**	47.19	51.5
+ General + AVID	63.17	**47.52**	**57.0**

Table 4: Qwen2.5-Omni-3B accuracy (in percent) after SFT on different dataset combinations. DO: DailyOmni; WS: WorldSense; AVID: inconsistency detection.



(a) Confusion matrices (b) Recall vs. FPR (c) Grounding quality

Figure 5: Complementary analyses of audio-visual inconsistency understanding: (a) segment-level 8-category confusion matrices; (b) segment-level recall versus false positive rate (upper-left is ideal); and (c) full-video grounding quantity versus quality (x-axis closer to zero and y-axis higher is ideal).

First, withholding video reduces segment detection accuracy from 61.3% to 53.3% for AVID-Qwen and from 69.7% to 55.8% for Gemini 3.1 Pro; vision-only performance is near chance. Second, we pass 100 semantically consistent clips through the same voice-cloned TTS pipeline while keeping the spoken content and speaker timbre aligned with the video. Relative to clean clips, false-positive rates increase by only 0.5 and 2.5 percentage points for Gemini 3.1 Pro and AVID-Qwen, respectively. AVID thus assesses models' audio-visual understanding and cannot be solved through artifact recognition alone.

Can AVID serve as a sub-task dataset for general omni-modal training? To validate the effectiveness of the AVID training split, we train Qwen2.5-Omni-3B on a 390K-sample general mixture and add 20K AVID segment examples (Table 4). The general mixture alone improves DailyOmni and WorldSense but leaves AVID detection near chance (51.5%). Adding AVID raises inconsistency detection to 57.0%, while general omni-modal performance remains largely unchanged. Moreover, without DAVE-specific tuning, AVID-Qwen improves over its Qwen3-Omni base from 37.3% to 47.9% overall accuracy on DAVE (Table 5). Together, these results show that the AVID training split provides targeted inconsistency supervision that transfers beyond its own evaluation format and serves as an effective sub-task dataset for training general omni-modal models. Full mixture and DAVE evaluation details are provided in the appendix.

5.4 In-depth Analysis

Fine-grained Classification. Figure 5a presents confusion matrices for eight-category classification. Gemini, MiMo-V2-Omni, and AVID-Qwen exhibit pronounced diagonal patterns across categories, indicating relatively consistent category discrimination. Other open-source models concentrate their correct predictions in one or two categories. For example, OLA primarily recognizes lip-sync and semantic inconsistencies, suggesting limited discrimination across the full taxonomy. **Segment-level Task.** Figure 5b shows that the Gemini models combine high recall with low false-positive rates. AVID-Qwen achieves a similar balance after fine-tuning. Other open-source models tend to overpredict or underpredict inconsistencies, resulting in elevated false-positive rates or reduced recall. **Full-video Task.** Figure 5c examines the relationship between the number of predicted intervals and localization precision. The former is measured by the difference between predicted and ground-truth interval counts, and the latter by the proportion of predictions exceeding an IoU threshold. Gemini and MiMo-V2-Omni predict fewer intervals than the ground truth but maintain relatively high precision. Open-source models generally exhibit lower precision even when their

Model	Segment				Full-Video					DAVE Transfer			
	Det.	Cla.	Reas.		Det.	Ground.		Reas.	Dense	Overall	Sync	Abs.	Discr.
	Acc	Acc	B	R	Acc	R@.5	mIoU	R	S_M	Acc	Acc	Acc	Acc
Qwen3-Omni	52.6	55.0	2.2	21.7	75.3	8.3	9.1	3.90	1.48	37.3	42.2	10.1	14.4
w/ Stage 1	**61.5**	**58.5**	**5.8**	**26.2**	62.5	4.6	5.1	2.64	1.25	–	–	–	–
w/ Stage 2	57.5	51.2	4.0	20.4	**76.4**	**31.6**	**31.7**	**14.6**	**6.11**	–	–	–	–
AVID-Qwen	**61.3**	**55.5**	**6.2**	**26.5**	**78.2**	**39.2**	**36.1**	**15.2**	**7.47**	**47.9**	**51.0**	**31.0**	**34.0**

Table 5: Ablation of the two-stage training strategy for AVID-Qwen and evaluation on DAVE without DAVE-specific fine-tuning. DAVE results are reported as accuracy percentages. Abs.: sound absence; Discr.: sound discrimination.

predicted interval counts are close to the ground truth. AVID-Qwen predicts more intervals and has lower proportion-based precision, but achieves an R@0.3 of 55.2%, compared with 30.9% for the next-best model (Table 2). These results indicate broader coverage of annotated inconsistency events, accompanied by a trade-off in prediction precision.

Qualitative Analysis. Figure 6 presents representative outputs from Qwen3-Omni and AVID-Qwen on both tasks. In these examples, the base model misses segment-level inconsistencies and produces full-video predictions with imprecise temporal boundaries. AVID-Qwen identifies the inconsistency categories, provides more specific explanations, and localizes multiple events with more accurate boundaries.

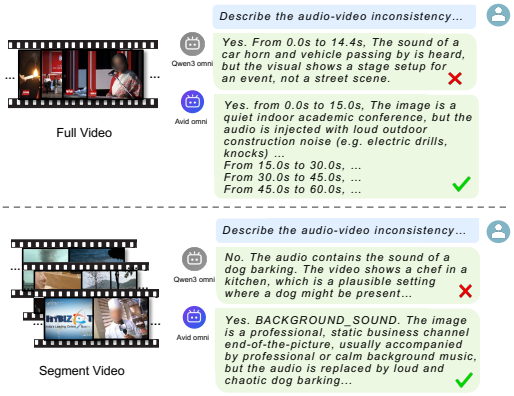


Figure 6: Qualitative comparison of Qwen3-Omni and AVID-Qwen.

5.5 Ablation Study

We assess the contribution of each training stage by comparing AVID-Qwen with variants trained using a single stage (Table 5). Stage 1 alone yields strong segment-level performance but achieves only 4.6% R@0.5 on full-video grounding. Stage 2 alone improves full-video grounding, while segment classification accuracy falls to 51.2% and reasoning performance declines relative to Stage 1. Combining both stages balances segment-level and full-video performance. These results suggest that the two stages provide complementary supervision: segment-level training develops fine-grained local perception, and full-video training extends this capability to temporal reasoning over multiple events. Classification accuracy improves only slightly from 55.0% to 55.5%, despite substantial gains in detection and reasoning. This difference suggests a potential misalignment between explanation generation and category prediction that merits further investigation.

6 Conclusion

We present AVID, a large-scale benchmark for systematic audio-visual inconsistency understanding in long-form videos, comprising 11.2K videos, 39.4K annotated inconsistency events, and 78.7K segment clips across eight physical and semantic categories. In contrast to controlled multiple-choice diagnostics such as DAVE, AVID emphasizes multi-event detection, temporal grounding, and free-form reasoning. Modality ablations and processed-consistent controls show that high scores require cross-modal comparison rather than merely recognizing manipulated audio, while general-mixture training demonstrates AVID's practical use as a complementary sub-task dataset. Evaluation of existing models shows that detecting inconsistencies does not ensure accurate temporal localization or faithful explanations. AVID-Qwen, trained in two stages, improves over its base model across detection, grounding, classification, and reasoning. These results establish AVID as both a diagnostic testbed and a targeted source of supervision for trustworthy omni-modal understanding.

LIMITATIONS

Audio-Only Manipulation. AVID modifies audio while preserving video frames and therefore does not cover visual or joint audio-visual manipulations. Video editing poses an additional challenge: producing a perceptible inconsistency while controlling synthesis artifacts that could provide detection shortcuts. Extending AVID to visual manipulations requires controls that distinguish cross-modal understanding from artifact recognition.

Language and Cultural Coverage. AVID uses English source videos and evaluation prompts. Its coverage of other languages and cultural contexts, where audio-visual consistency cues may differ, remains limited.

Context-Dependent Inconsistency. Background sounds and emotional music may appear inconsistent with visual content while serving artistic purposes such as irony or satire. Recognizing such intent without incorrectly labeling deliberate contrasts as inconsistencies requires a more advanced level of contextual understanding. Extending AVID to evaluate this capability is an important direction for future work.

AI USE STATEMENT

We used AI assistants to aid manuscript drafting, translation, and language polishing, and to support literature retrieval and discovery. We also used generative models to construct synthetic data: Gemini 3.1 Pro planned audio-visual inconsistency injections, and Qwen3-TTS synthesized replacement speech. The construction pipeline includes human review, as described in Section 3.4. The authors take responsibility for the final manuscript, references, and research artifacts, including all AI-assisted content.

ETHICS STATEMENT

AVID derives from the publicly released LongVALE collection (Geng et al., 2025), whose source videos originate from YouTube. We release derived annotations and the audio manipulation pipeline without redistributing raw videos. Users must comply with source-platform terms and takedown requests. Identity modification uses acoustic transformations, while voice cloning uses only the speaker reference within the same segment. Manipulated segments are labeled as synthetic and include provenance metadata. AVID is intended for research on omni-modal training and evaluation; deceptive or non-consensual applications are prohibited, and the release terms restrict generated-audio redistribution to research contexts.

Annotation was conducted by research-group members with informed consent and compensation at local fair-wage standards. Annotators were briefed on the content and its intended use and could skip samples or withdraw at any time. The study was determined to be exempt from formal institutional ethics review. Annotation and human evaluation protocols are detailed in Appendices A.4 and E.

REPRODUCIBILITY STATEMENT

Appendix A documents data construction, injection prompts, and quality control. Appendix B provides model and training configurations, and Appendix C specifies evaluation prompts and metrics. Additional training and transfer protocols appear in Appendices D.1 and D.2; the human evaluation protocol is described in Appendix E.

REFERENCES

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do You Really Mean That? Content Driven Audio-Visual Deepfake Dataset and Multimodal Method for Temporal Forgery Localization. In *2022 International Conference on Digital Image Computing: Techniques and*

Applications (DICTA), pp. 1–10, Sydney, Australia, 2022. IEEE. ISBN 978-1-6654-5642-5. doi: 10.1109/DICTA56598.2022.10034605. URL https://ieeexplore.ieee.org/document/10034605/.

Zhixi Cai, Kartik Kuckreja, Shreya Ghosh, Akanksha Chuchra, Muhammad Haris Khan, Usman Tariq, Tom Gedeon, and Abhinav Dhall. AV-Deepfake1M++: A Large-Scale Audio-Visual Deepfake Benchmark with Real-World Perturbations. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, 2025. doi: 10.1145/3746027.3761979.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70M: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Alexandre Défossez. Hybrid spectrogram and waveform source separation. In *Proceedings of the ISMIR 2021 Music Demixing Workshop*, 2021.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. SODA: Story oriented dense video captioning evaluation framework. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 517–531, 2020.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5267–5275, 2017.

Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, Jinwen Luo, Weibo Gu, Zexuan Li, Xiaojing Zhang, Yangyu Tao, Han Hu, Di Wang, and Ying Shan. ARC-Hunyuan-Video-7B: Structured video comprehension of real-world shorts, 2025. URL https://arxiv.org/abs/2507.20939. arXiv:2507.20939.

Gemini Team, Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. 2025. URL https://arxiv.org/abs/2507.06261. arXiv:2507.06261.

Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. LongVALE: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Arushi Goel, Karan Sapra, Matthieu Le, Rafael Valle, Andrew Tao, and Bryan Catanzaro. OMCAT: Omni context aware transformer, 2024. URL https://arxiv.org/abs/2410.12109. arXiv:2410.12109.

Google DeepMind. Gemma 4, 2026. URL https://deepmind.google/models/gemma/gemma-4/. Model card. Accessed: 2026-05-15.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. TrustLLM: Trustworthiness in large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. ACAV100M: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

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

Weifeng Liu, Tianyi She, Jiawei Liu, Boheng Li, Dongyu Yao, Ziyou Liang, and Run Wang. Lips Are Lying: Spotting the Temporal Inconsistency between Audio and Visual in Lip-Syncing DeepFakes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b. URL https://arxiv.org/abs/2401.15668.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model, 2025. URL https://arxiv.org/abs/2502.04328. arXiv:2502.04328.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024. doi: 10.1109/TASLP.2024.3419446.

Trevine Oorloff, Surya Koppisetti, Nicolo Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2406.02951.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, 2002.

Gorjan Radevski, Teodora Popordanoska, Matthew B. Blaschko, and Tinne Tuytelaars. DAVE: Diagnostic benchmark for audio visual evaluation. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=4ZAX1NT0ms.

Silero Team. Silero VAD: Pre-trained enterprise-grade voice activity detector, 2021. URL https://github.com/snakers4/silero-vad. GitHub repository.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-SALMONN: Speech-enhanced audio-visual large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2406.15704.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. InternVid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiaomi AI Lab. MiMo: Unlocking the reasoning potential of language model – from pretraining to posttraining, 2025. URL http://arxiv.org/abs/2505.07608. arXiv:2505.07608 [cs].

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni Technical Report, 2025. URL http://arxiv.org/abs/2509.17765. arXiv:2509.17765.

Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake. *IEEE Transactions on Information Forensics and Security*, 18:2015–2029, 2023. ISSN 1556-6013, 1556-6021. doi: 10.1109/TIFS.2023.3262148. URL https://ieeexplore.ieee.org/document/10081373/.

Qilang Ye, Wei Zeng, Meng Liu, Jie Zhang, Yupeng Hu, Zitong Yu, and Yu Zhou. When eyes and ears disagree: Can mllms discern audio-visual confusion? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11955–11963, 2026.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations(EMNLP)*, pp. 543–553, 2023.

A Benchmark Construction Details

A.1 Class-Aware Temporal Segmentation

Each video is segmented into class-aware units using a four-step pipeline. **Step 1 (VAD):** Silero VAD detects speech activity on 16 kHz mono audio (threshold 0.5, min speech duration 250 ms, min silence duration 300 ms). Adjacent speech segments within a 0.3 s gap are merged. **Step 2 (Face Detection):** OpenCV Haar Cascade (haarcascade_frontalface_default, scaleFactor = 1.1, minNeighbors = 5, minSize = 30×30) is applied at 5 fps. A face-present segment ends after 5 consecutive no-face frames ($\approx$1 s). **Step 3 (Classification):** Each time interval is labeled by combining VAD and face detection results: no speech $\rightarrow$ Class 3 (Scenic), speech + face $\rightarrow$ Class 1 (Active Speaker), speech + no face $\rightarrow$ Class 2 (Voiceover). **Step 4 (Temporal Fusion):** Adjacent segments with the same class label and a gap < 0.1 s are merged. Class 3 segments shorter than 1.0 s are absorbed into their neighbors to remove brief speech pauses. All timestamps are rounded to 10 ms.

Speech?	Active Face?	Class	#Segments	Share (%)
Yes	Yes	Class 1 (Active Speaker)	32,908	41.80
Yes	No	Class 2 (Voiceover)	18,096	22.99
No	N/A	Class 3 (Scenic)	27,712	35.20

Table 6: Decision logic and per-class segment statistics for temporal segment classification.

A.2 STRATEGY PLANNING

The Strategy Planning Agent (Gemini 3.1 Pro) receives the temporal segmentation table and decides which segments to inject and which inconsistency type to apply. The full prompt is shown in Figure 7. The agent outputs a structured JSON plan containing segment boundaries, injection type, parameters, and a natural-language reasoning for each selected segment. The prompt is iteratively refined through human feedback as described in §3.4.

```
# Strategy planning prompt
```

You are a professional expert in constructing audio-visual inconsistency data. Please select appropriate injection strategies for a video based on its temporal segmentation information.
The input video contains the following information: total duration {`duration:.1f`} seconds, {`len(valid_segments)`} valid segments, and a temporal segmentation table. Please use the provided time ranges directly and do not estimate them yourself.

Time range (duration) | category
{`segments_text`}

The segments are divided into three categories: `class_1_active_speaker` (speech with visible speaker), `class_2_voiceover` (speech without visible speaker), and `class_3_scenic` (no speech, only background sound). The corresponding optional injection types are:

- `class_1_active_speaker`: `VOICE_IDENTITY`, `VOLUME_FLUCTUATION`, `LIP_SYNC`, `TEMPORAL_SHIFT`
- `class_2_voiceover`: `BACKGROUND_CONFLICT`, `SEMANTIC_DIVERGENCE`
- `class_3_scenic`: `EMOTION_MISMATCH`, `BACKGROUND_SOUND`

Please analyze the video and generate an injection plan under the following requirements:

1. **Segment selection.** Select appropriate segments for injection. Each selected segment must be at least 5 seconds and at most 20 seconds long. If a segment exceeds 20 seconds, subdivide it into smaller non-overlapping sub-segments, each treated as an independent injection plan.
2. **Balance and preference.** Construct {`min_injections`}–{`max_injections`} injections depending on video length. Maintain class balance whenever possible by including at least one segment from each available class. Diversify injection types across selected segments. Prefer `TEMPORAL_SHIFT` and `LIP_SYNC`, and use `EMOTION_MISMATCH` and `BACKGROUND_SOUND` less frequently.
3. **Inconsistency analysis.** For each selected segment, assign one valid injection type and provide a detailed explanation including visual description, audio description, inconsistency point, a fine-grained `dataset_sub_category`, and reasoning.
4. **Output format.** Return the result in strict JSON format:

```
{
  "injection_plans": [{
    "start": "...",
    "end": "...",
    "class_label": "...",
    "injection_type": "...",
    "injection_params": {...},
    "inconsistency_understanding": {
      "visual_description": "...",
      "audio_description": "...",
      "inconsistency_point": "..."
    },
    "dataset_sub_category": "...",
    "reasoning": "..."
  }],
  "summary": "..."
}
```

Important constraints: selected segments must not overlap; `injection_type` must be chosen from the predefined list; and `injection_params` must follow the corresponding format. In particular, `TEMPORAL_SHIFT` uses {`"shift_seconds": 1.0`} (0.5–3.0 seconds; positive = audio lags, negative = audio leads); `LIP_SYNC` uses {`"text": "...", "voice_type": "young_female"`}; `SEMANTIC_DIVERGENCE` uses {`"contradictory_text": "...", "voice_type": "young_female"`}, where text length should match clip duration (5–10 s: 15–25 words; 10–15 s: 25–35; 15–20 s: 35–50); `VOICE_IDENTITY` uses {`"target_voice": "Male"`}; `VOLUME_FLUCTUATION` uses {`"direction": "away"`} or {`"direction": "toward"`}; `BACKGROUND_CONFLICT` and `BACKGROUND_SOUND` use {`"bg_sound_type": "city_traffic"`}; and `EMOTION_MISMATCH` uses {`"emotion": "happy"`}.

Figure 7: Prompt for strategy planning agent.

A.3 INJECTOR DETAILS

Each injector follows a unified interface $\mathcal{I}_k : (V, A, M) \to (V', A', \Delta, \Theta_k)$, where V/A are original video/audio streams, M is strategy metadata, Δ is the annotation delta, and Θ_k is injector-specific configuration. Tables 7 and 8 summarize the five injectors and their class-conditional mapping. All injectors operate on the audio stream only; the original video frames, resolution, and duration are preserved. Audio is encoded as AAC at 192 kbps, 48 kHz stereo.

#	Category	Core Mechanism	Key Tools
1	TemporalInjector (Temporal Shift)	Audio time-axis offset ($\pm$0.5–3.0 s)	FFmpeg (PCM extraction + temporal remapping)
2	SemanticInjector (Semantic Contradiction)	Replace speech with contradictory TTS while preserving background	Demucs (Défossez, 2021) + Qwen3-TTS + FFmpeg (amix)
3	IdentityInjector (Vocal Identity Conflict)	Modify pitch/formant/brightness for identity mismatch	FFmpeg (asetrate, atempo, equalizer, tremolo)
4	SpatialInjector (Spatial Acoustics)	Distance-based volume attenuation (near/far field)	FFmpeg (volume envelope)
5	BackgroundInjector (Environmental/Affective Conflict)	Replace background audio with scene-contradictory ambience or affective music	Demucs + Background Library + FFmpeg

Table 7: Overview of the five injection techniques in AVID.

Segment Class	Inconsistency Category	Injector	Description
Class 1 (Active Speaker)	VOICE_IDENTITY	Identity	Timbre/identity conflict (age/gender/voice-profile shift)
	VOLUME_FLUCTUATION	Spatial	Loudness changes without corresponding visual movement
	LIP_SYNC	Semantic	Mismatch between generated speech and lip motion
	TEMPORAL_SHIFT	Temporal	Audio leads/lags video by 0.5–3.0 s
Class 2 (Voiceover)	BACKGROUND_CONFLICT	Background	Narration/music conflicts with visual scene context
	SEMANTIC_DIVERGENCE	Semantic	Spoken content contradicts visual semantics
Class 3 (Scenic)	EMOTION_MISMATCH	Background	Background mood conflicts with scene affect
	BACKGROUND_SOUND	Background	Background sound conflicts with scene environment

Table 8: Class-conditional mapping between segment classes and applicable inconsistency categories.

Temporal Shift applies an audio time-axis offset $|\delta t| \in [0.5, 3.0]$ s via FFmpeg PCM extraction. Positive values delay audio (prepend silence, truncate end); negative values advance it (skip beginning, append silence).

Semantic Divergence replaces original speech with contradictory TTS content: (1) separate vocals from background using Demucs (htdemucs, two-stems); (2) generate contradictory speech via Qwen3-TTS (12Hz-1.7B-Base) with voice cloning from the original segment; (3) match duration using FFmpeg `atempo` $\in [0.7, 1.3]$; (4) mix new vocals with preserved background (volume ratio 1.0 : 0.8). Text length scales with segment duration (15–25 words for 5–10 s, 25–35 for 10–15 s, 35–50 for 15–20 s).

Identity Modification modifies vocal characteristics via FFmpeg filters (`asetrate` for pitch ±5–10 semitones, `equalizer` for formant scaling 0.75–1.25, optional `tremolo` f=4.5 Hz for age effects). Seven presets are defined: Female, Female_Young, Female_Old, Male, Male_Deep, Child, Elder.

Spatial Mismatch applies volume envelope modulation from 0.01 (far-field) to 1.0 (near-field), with a smooth fade over 50% of the segment duration. Two direction modes: "away" ($1.0 \rightarrow 0.01$) and "toward" ($0.01 \rightarrow 1.0$).

Background Replacement separates foreground vocals via Demucs, then re-mixes them with replacement audio from a library of 20 sound categories (6 urban, 5 nature, 2 animal, 2 machinery, 5 music-emotion). Volume ratio: 1.0 for vocals, 0.5–0.6 for replacement background. Duration is matched via truncation or looping.

A.4 Annotation Quality

Inter-annotator agreement is measured at two stages of the construction pipeline. All annotators are members of the research group with backgrounds in computer science and multimodal learning. At the strategy planning stage, 3 annotators independently evaluated 200 sampled segments on injection suitability and category appropriateness (Cohen's $\kappa = 0.78$). Annotators were instructed to reject strategies where (1) the selected segment contained ambiguous or insufficient audio-visual content to support a perceptible inconsistency, (2) the chosen inconsistency type was incompatible with the segment class, or (3) the injection parameters would produce imperceptible or unrealistic modifications. At the post-construction quality control stage, the same 3 annotators verified 300 videos (150 inconsistent, 150 consistent) on detection correctness, category label, and temporal boundary accuracy (within 1 s tolerance): detection $\kappa = 0.85$, classification $\kappa = 0.81$, boundary $\kappa = 0.79$. Annotators rejected videos where (1) the injected inconsistency was imperceptible or ambiguous to a naive viewer, (2) the category label did not match the perceivable inconsistency, or (3) injection artifacts introduced unintended visual or audio distortion.

The 300-video subset above is used to measure agreement; final quality control on the inconsistent test partition is exhaustive. Annotators inspect all 1,250 constructed inconsistent test videos and remove 79 poor or ambiguous cases (6.3%), leaving 1,171. Consistent videos are consistent by construction and are sampled to preserve the 3:1 inconsistent-to-consistent test ratio. The train/test partition is disjoint at the source-video level: 9,639/1,561 full videos are split first, after which every segment inherits its parent video's split, producing 68,088/10,634 train/test segments. Thus, no source video or any of its derived segments appears in both partitions.

B Model and Training Details

B.1 Model Overview

Qwen3-Omni is an omni-modal large language model with approximately 30 billion total parameters (3 billion activated). It consists of four components: a **Vision Encoder** that converts input video frames into visual tokens, an **Audio Encoder** that extracts acoustic features from the video's audio stream, an **Aligner Module** that bridges vision encoder outputs to the LLM embedding space, and a decoder-only **LLM Backbone** that performs multimodal reasoning over concatenated visual, audio, and text tokens. As illustrated in Figure 4, the vision encoder and aligner are frozen, while the audio encoder is trainable and the LLM backbone is adapted via LoRA.

B.2 Input Preprocessing

Input preprocessing follows the default Qwen3-Omni configuration. Videos are sampled at 12 frames per second. Each frame is resized to a maximum of 50,176 pixels while preserving aspect ratio, with visual token sequences capped at 1,003,520 pixels. Audio is extracted directly from the video file using the decord reader at the model's default sampling rate. The maximum token sequence length is set to 8,192. The model accepts a conversation format with interleaved video and text; the `<video>` token triggers automatic frame sampling and audio extraction.

B.3 Training Hyper-Parameters

We provide the complete training hyper-parameters for both fine-tuning stages. Table 9 summarizes the configuration.

Parameter	Stage 1 (Segment)	Stage 2 (Full-Video)
Base Model	Qwen3-Omni-30B-A3B-Instruct	
Training Type	LoRA	
LoRA Rank	8	
LoRA Alpha	32	
LoRA Target Modules	all-linear	
Frozen Modules	Vision Encoder, Aligner	
Learning Rate	1e-4	
Optimizer	AdamW (bfloat16)	
Warmup Ratio	0.05	
Per Device Batch Size	2	
Gradient Accumulation Steps	4	
Effective Batch Size	48 (6 GPUs × 2 × 4)	
Training Epochs	2	10
Max Sequence Length	8192	
Video Sampling Rate (fps)	12	
Video Max Pixels	50,176	
Hardware	6 × NVIDIA A100 80GB	
Distributed Training	DeepSpeed ZeRO-3	

Table 9: Training hyper-parameters for AVID-Qwen two-stage fine-tuning.

Training is conducted with the SWIFT framework using DeepSpeed ZeRO-3 across 6 A100 80GB GPUs with gradient checkpointing enabled.

C Evaluation Details

C.1 Compared Models and Inference Settings

Table 10 summarizes the model versions and test-time configurations. Closed-source models (Gemini 2.5/3.1 Pro, MiMo-V2-Omni) are accessed via API with audio embedded in the video file through base64 encoding. Open-source models (Qwen3-Omni, OLA, ARC-Hunyuan-Video, Video-SALMONN 2, Gemma4 E4B) are deployed locally on NVIDIA A100 80GB GPUs using HuggingFace checkpoints. Models use temperature-based sampling with the decoding settings listed in Table 10. Video sampling is model-dependent: Gemini models use automatic frame selection while local models use fixed FPS (8–16 fps). Audio is extracted directly from video files using FFmpeg; local models process audio at their native sample rate. Evaluation follows the task sequence in Figure 4: detection is followed by category classification and reasoning for segments, or temporal grounding and dense reasoning for full videos. The corresponding prompts are provided in Figures 8 and 9. API timeout is set to 600 seconds with no retries to ensure reproducibility; malformed outputs are parsed using regex to extract partial information when possible.

C.2 Evaluation Protocol

All metrics are reported in percentage scale ($\times 100$). For each model, we use the fixed decoding configuration in Table 10 and evaluate each sample once. We evaluate on two complementary tasks: (i) segment-level detection, classification, and reasoning (Figure 8); and (ii) full-video detection with temporal grounding and dense reasoning (Figure 9). Both prompts include the complete 8-category inconsistency taxonomy and enforce structured output format. Classification, grounding, and reasoning metrics are computed only on correctly detected samples (True Positives).

Model	Provider/API	Version/Checkpoint
Gemini 2.5 Pro	Google Vertex AI	gemini-2.5-pro-preview-0506
Gemini 3.1 Pro	Google Vertex AI	gemini-3.1-pro-preview
ARC-Hunyuan-Video	Local (HuggingFace)	TencentARC/ARC-Hunyuan-Video-7B
Qwen3-Omni	Local (HuggingFace)	Qwen3-Omni-30B-A3B-Instruct
OLA	Local	Ola-main/inference
Video-SALMONN 2-7B	Local	Video-SALMONN-2-7B
Video-SALMONN 2-72B	Local	Video-SALMONN-2-72B
MiMo-V2-Omni	OpenRouter	xiaomi/mimo-v2-omni
Gemma4 E4B	Local (HuggingFace)	google/gemma-4-e4b-it
AVID-Qwen (w/ Stage 1)	Local (Qwen3-Omni)	Qwen3-Omni + LoRA Stage 1
AVID-Qwen	Local (Qwen3-Omni)	Qwen3-Omni + LoRA Stage 1+2

Model	Context	Temp.	Top-p	Max output tokens	Video sampling	Audio input
Gemini 2.5 Pro	1M	0.3	-	4096	auto	Yes
Gemini 3.1 Pro	1M	0.3	-	4096	auto	Yes
ARC-Hunyuan-Video	32K	0.7	0.9	2048	8fps	Yes
Qwen3-Omni	8K	0.3	-	2048	12fps	Yes
OLA	8K	0.3	-	2048	12fps	Yes
Video-SALMONN 2-7B	8K	0.3	-	2048	–	Yes
Video-SALMONN 2-72B	8K	0.3	-	2048	–	Yes
MiMo-V2-Omni	32K	0.3	-	4096	auto	Yes
Gemma4 E4B	8K	0.3	-	2048	12fps	30s max
AVID-Qwen (w/ Stage 1)	8K	0.3	-	2048	12fps	Yes
AVID-Qwen	8K	0.3	-	2048	12fps	Yes

Table 10: Model identifiers (top) and inference settings (bottom). Video sampling is in frames per second; "auto" denotes automatic sampling. Video-SALMONN 2 uses 16 kHz audio; "– indicates an unspecified video sampling rate. Gemma4 E4B has a 30-second audio limit.

C.3 Metric Definitions

For a sample i, let $y_i \in \{0, 1\}$ denote the binary inconsistency label, $\hat{y}_i$ the model prediction, c_i and $\hat{c}_i$ the ground-truth and predicted inconsistency categories, and $\hat{r}$/r the generated/reference explanations. For temporal grounding, let the predicted interval be $\hat{g}_i = [\hat{t}_s, \hat{t}_e]$ and the ground-truth interval be $g_i = [t_s, t_e]$.

Binary Detection.

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i), \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{2}$$

$$\text{FPR} = \frac{FP}{FP + TN}. \tag{3}$$

Category Classification. Let $\mathcal{T} = \{i : y_i = \hat{y}_i = 1\}$ be the correctly detected inconsistent samples. Consistent with Table 2, classification accuracy is conditioned on this set:

$$\text{Acc}_{\text{type}} = \frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}}\mathbb{I}(\hat{c}_i = c_i). \tag{4}$$

```
# Prompt for audio-visual inconsistency understanding
```

Please carefully watch this video, analyze its audio and visual content, and determine whether there is audio-visual inconsistency.

Output format
Please return in the following exact format:

1. **Is there inconsistency:** (Yes/No) – Judge independently based on video content.
2. **Inconsistency category:** (If Yes, choose one from the following 8 categories. If No, fill `No`.)
3. **Inconsistency point description:** (If inconsistency exists, describe the specific inconsistency between visual and audio; if No, fill `Audio and video are consistent`.)

8 Inconsistency Categories (must choose from these)

Class 1 (Active Speaker – interview/dialogue, person speaking in video)

- `TEMPORAL_SHIFT`: Temporal offset, audio leads or lags behind video by 0.5–3s (e.g., the mouth moves but the sound is delayed by 0.5s).
- `LIP_SYNC`: Lip-sync mismatch, TTS-generated voice does not match lip movement (e.g., the lips continue moving after the spoken phrase has ended).
- `VOICE_IDENTITY`: Voice identity conflict, the speaker's voice changes abruptly (e.g., the video shows an elderly person but the voice is a child's).
- `VOLUME_FLUCTUATION`: Volume conflict, the person is still but the volume fluctuates (e.g., the person stands still but the voice sounds as if it is moving closer and farther away).

Class 2 (Voiceover – narration, no speaker in video)

- `SEMANTIC_DIVERGENCE`: Semantic inconsistency, TTS text contradicts video content (e.g., the video shows food but the narration talks about phones).
- `BACKGROUND_CONFLICT`: Background sound conflict, narration or music contradicts the video scene (e.g., the video shows an office but the background contains bar music).

Class 3 (Scenic – scenery/scene, no human voice)

- `EMOTION_MISMATCH`: Background music emotion mismatch, the video is a sad scene but has happy music (e.g., the video shows a funeral but the music is joyful).
- `BACKGROUND_SOUND`: Background sound conflict, the video is scene A but the audio is scene B (e.g., the video shows a forest but has city traffic sounds; or the video shows chopping but no chopping sound).

Task sequence

1. **Question 1:** Is there audio-visual inconsistency? Answer `Yes` or `No` only.
2. **Question 2:** What is the inconsistency category? Choose from: `EMOTION_MISMATCH`, `BACKGROUND_SOUND`, `SEMANTIC_DIVERGENCE`, `LIP_SYNC`, `TEMPORAL_SHIFT`, `VOLUME_FLUCTUATION`, `VOICE_IDENTITY`, `BACKGROUND_CONFLICT`. If consistent, answer `No`.
3. **Question 3:** Please explain your reasoning process for the above judgment. Describe what you observed in the video and what you heard in the audio.

Figure 8: Segment-level evaluation prompt for audio-visual inconsistency detection, classification, and reasoning.

Text Reasoning. BLEU-4, ROUGE-L, and METEOR are computed between $\hat{r}$ and r under standardized tokenization and lower-casing, with the official maximum-over-references protocol when multiple references exist.

$$\text{BLEU-4} = \text{BP} \cdot \exp\left(\sum_{n=1}^{4} \tfrac{1}{4} \log p_n\right), \tag{5}$$

Model	Grounding							
	R@0.3		R@0.5		R@0.7		mIoU	
	R_d	P_d	R_d	P_d	R_d	P_d	R_d	P_d
Arc-Hunyuan-Video	6.7	6.7	4.2	4.2	1.9	1.9	4.4	4.4
OLA	8.9	11.3	4.4	5.6	1.7	2.2	5.5	6.9
SALMONN 2 -7B	0.0	25.0	0.0	25.0	0.0	0.0	0.0	13.8
Qwen3-Omni	12.3	9.3	8.3	6.3	4.9	3.7	9.1	6.9
MiMo-V2-Omni	12.2	23.3	11.6	22.2	9.9	18.9	10.2	19.5
Gemini 2.5 Pro	27.7	24.0	25.6	22.2	23.5	20.4	24.0	20.9
Gemini 3.1 Pro	29.5	25.6	26.8	23.3	24.2	21.0	25.0	21.7
AVID-Qwen	**53.5**	**43.5**	**38.0**	**30.9**	**18.5**	**15.1**	**35.0**	**28.4**

Model	Reasoning						Dense	
	B		R		M		S_M	
	R_d	P_d	R_d	P_d	R_d	P_d	R_d	P_d
Arc-Hunyuan-Video	0.01	0.01	1.06	1.06	0.50	0.50	0.39	0.39
OLA	0.11	0.13	2.27	2.87	1.21	1.53	0.75	0.95
SALMONN 2 -7B	0.00	0.16	0.01	6.45	0.00	4.95	0.00	0.38
Qwen3-Omni	0.25	0.19	3.89	2.94	2.48	1.87	1.48	1.11
MiMo-V2-Omni	0.28	0.53	2.87	5.50	2.65	5.08	2.80	5.37
Gemini 2.5 Pro	0.63	0.55	6.39	5.55	5.43	4.71	6.08	5.28
Gemini 3.1 Pro	0.55	0.48	6.71	5.83	4.88	4.24	5.56	4.83
AVID-Qwen	**2.65**	**2.15**	**14.74**	**11.98**	**14.55**	**11.82**	**7.25**	**5.89**

Table 11: Full-video grounding, reasoning, and dense understanding metrics with recall- and precision-denominator variants. R_d: recall-denominator ($\div$(TP+FN)); P_d: precision-denominator ($\div$(TP+FP)).

$$\mathrm{BP} = \begin{cases} 1 & \text{if } |\hat{r}| > |r| \\ \exp(1 - |r|/|\hat{r}|) & \text{otherwise} \end{cases}, \tag{6}$$

where p_n is the modified n-gram precision.

$$\text{ROUGE-L} = \frac{(1+\beta^2) R_{\text{lcs}} P_{\text{lcs}}}{R_{\text{lcs}} + \beta^2 P_{\text{lcs}}}, \tag{7}$$

$$R_{\text{lcs}} = \frac{\text{LCS}(\hat{r}, r)}{|r|}, \tag{8}$$

$$P_{\text{lcs}} = \frac{\text{LCS}(\hat{r}, r)}{|\hat{r}|}, \tag{9}$$

where $\beta = 1.2$.

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \gamma \cdot \text{frag}^{\theta}), \tag{10}$$

$$F_{\text{mean}} = \frac{10 P_m R_m}{R_m + 9 P_m}, \tag{11}$$

where $P_m = |m|/|\hat{r}|$, $R_m = |m|/|r|$, $|m|$ is the number of matched unigrams (including stems, synonyms, and paraphrases), frag = #chunks$/|m|$, $\gamma = 0.5$, $\theta = 3$.

Temporal Grounding.

$$\text{IoU}(\hat{g}_i, g_i) = \frac{|\hat{g}_i \cap g_i|}{|\hat{g}_i \cup g_i|}, \tag{12}$$

$$\text{mIoU} = \frac{1}{N_g} \sum_{i=1}^{N_g} \text{IoU}(\hat{g}_i, g_i), \tag{13}$$

```
# Full-video inconsistency detection prompt
```

Please carefully watch this full video (not a short segment), analyze its audio and visual content, and determine whether there is audio-visual inconsistency.

8 Inconsistency Categories (must choose from these)

Class 1 (Active Speaker – interview/dialogue, person speaking in video)

- `TEMPORAL_SHIFT`: Temporal offset, audio leads or lags behind video by 0.5–3s (e.g., see mouth moving but sound is delayed by 0.5s).
- `LIP_SYNC`: Lip-sync mismatch, TTS-generated voice does not match lip movement (e.g., the lips continue moving after the spoken phrase has ended).
- `VOICE_IDENTITY`: Voice identity conflict, speaker's voice changes abruptly (e.g., video shows an elderly person but voice is a child's).
- `VOLUME_FLUCTUATION`: Volume conflict, person is still but volume fluctuates (e.g., person stands still but voice sounds like it is moving closer and further).

Class 2 (Voiceover – narration, no speaker in video)

- `SEMANTIC_DIVERGENCE`: Semantic inconsistency, TTS text contradicts video content (e.g., video shows food but narration talks about phones).
- `BACKGROUND_CONFLICT`: Background sound conflict, narration/music contradicts video scene (e.g., video shows office but background has bar music).

Class 3 (Scenic – scenery/scene, no human voice)

- `EMOTION_MISMATCH`: Background music emotion mismatch, video is sad scene but has happy music (e.g., video shows funeral but music is joyful).
- `BACKGROUND_SOUND`: Background sound conflict, video is scene A but audio is scene B (e.g., video shows forest but has city traffic sounds; video shows chopping but no chopping sound).

System instruction

Please answer in the following exact format (each item on a new line):

1. **Is there any inconsistency in this video?** Answer: `Yes` or `No`.
2. **If Yes, for each inconsistency, provide:** `from X.Xs to Y.Ys, reasoning` – each on a separate line. If No, write `N/A`.

Example output format

```
Yes
from 0.0s to 7.8s, The sad background music creates
emotional conflict with the lively visual tone
from 15.5s to 20.3s, Rain is audible but no rain is
shown in the indoor scene
```

Figure 9: Full-video evaluation prompt for audio-visual inconsistency detection, temporal grounding, and dense reasoning.

$$\mathrm{R@1}(\alpha) = \frac{1}{N_g} \sum_{i=1}^{N_g} \mathbb{I}(\mathrm{IoU}(\hat{g}_i, g_i) \geq \alpha)\,, \tag{14}$$

$$\alpha \in \{0.3, 0.5, 0.7\}. \tag{15}$$

Dense Full-Video (SODA-m). Given predicted events $\hat{\mathcal{E}} = \{(\hat{g}_j, \hat{c}_j)\}_{j=1}^{M}$ and ground-truth events $\mathcal{E} = \{(g_i, c_i)\}_{i=1}^{N}$, SODA-m first computes an optimal bipartite matching via the Hungarian algorithm $\pi^* = \arg\max_\pi \sum_i \mathrm{IoU}(g_i, \hat{g}_{\pi(i)})$. The per-pair score is $s_i = \mathrm{IoU}(g_i, \hat{g}_{\pi^*(i)}) \cdot S(c_i, \hat{c}_{\pi^*(i)})$, where S is a text similarity function (METEOR). The final score aggregates over all events:

$$\text{SODA-m} = \frac{\sum_{i=1}^{|\mathcal{M}|} s_i}{N + (M - |\mathcal{M}|)}, \tag{16}$$

Model	BG_CONFLICT			BG_SOUND			EMOTION			LIP_SYNC		
	B	R	M	B	R	M	B	R	M	B	R	M
Hunyuan	1.3	16.4	17.9	1.4	16.5	17.4	1.3	15.8	16.9	1.7	17.1	16.9
OLA	1.1	14.4	11.8	1.0	15.0	11.9	1.2	16.4	11.9	1.2	16.8	12.5
SALMONN-72B	0.5	9.6	9.5	1.5	16.4	20.9	0.8	13.3	15.7	2.9	20.1	18.3
Gemma4 E4B	-	-	-	-	-	-	1.7	12.7	13.9	2.1	21.0	22.0
Qwen3-Omni	2.2	21.5	24.0	2.1	22.1	24.3	1.4	19.0	23.4	2.1	19.7	20.8
MiMo-V2	1.5	17.5	17.6	2.4	22.2	22.5	1.1	19.6	17.3	1.8	19.9	20.3
Gemini 2.5	2.0	20.6	22.1	2.3	21.4	22.2	1.7	19.1	22.0	2.1	19.5	20.1
Gemini 3.1	2.2	21.6	23.0	2.6	22.2	22.9	1.7	19.1	22.0	2.4	19.7	20.8
AVID-Qwen	**6.0**	**28.0**	**27.8**	**7.4**	**30.1**	**29.1**	**4.9**	**25.3**	**25.4**	**4.9**	**25.0**	**22.9**

Model	SEMANTIC			TEMPORAL			VOICE_ID			VOLUME		
	B	R	M	B	R	M	B	R	M	B	R	M
Hunyuan	1.4	16.8	16.7	1.4	17.9	17.4	1.2	14.9	16.6	1.2	15.8	15.9
OLA	1.9	19.9	15.1	0.9	16.8	11.4	1.8	15.3	12.0	1.4	14.0	11.2
SALMONN-72B	2.6	20.3	21.1	1.1	16.7	14.3	1.3	11.7	10.7	1.2	14.8	14.1
Gemma4 E4B	1.8	23.6	19.4	-	-	-	-	-	-	-	-	-
Qwen3-Omni	2.6	23.5	25.5	1.5	18.2	18.7	2.4	20.7	22.3	1.7	17.7	18.8
MiMo-V2	2.9	24.8	24.6	-	-	-	4.5	**25.8**	24.2	-	-	-
Gemini 2.5	2.5	22.7	23.2	1.4	18.4	20.0	2.9	21.6	23.7	2.2	18.5	**20.9**
Gemini 3.1	2.6	23.6	24.5	1.8	19.9	20.6	3.8	23.4	**25.5**	2.8	19.2	20.1
AVID-Qwen	**5.1**	**27.6**	**26.6**	**5.0**	**24.6**	**22.7**	**5.0**	23.5	23.9	**3.0**	**21.2**	18.7

Table 12: Category-wise segment-level reasoning scores on correctly detected samples. B/R/M: BLEU-4, ROUGE-L, and METEOR. The best score for each metric within each category is in **bold**.

where $\mathcal{M}$ is the set of matched pairs with IoU > 0 and $(M - |\mathcal{M}|)$ counts false positive predictions.

C.4 VIDEO-SALMONN 2 DETECTION BEHAVIOR

Video-SALMONN 2 (both 7B and 72B) predicts "No inconsistency" for nearly all samples, yielding ≈25% full-video detection accuracy (matching the consistent-class proportion under the 3:1 ratio). We verified this is not an implementation artifact: (1) prompting the model to describe video content without asking about inconsistency produces accurate audio-visual descriptions; (2) varying temperature (0.3, 0.7, 1.0) and enabling `do_sample` does not change the prediction; (3) raw outputs correctly parse as "No."

C.5 IN-DEPTH ANALYSIS DETAILS

This subsection provides the formal definitions for the analyses in §5.4.

Recall vs. False Positive Rate (Figure 5b). For segment-level binary detection, each model produces a prediction $\hat{y}_i \in \{0, 1\}$ for every sample. Recall and False Positive Rate are computed as:

$$\text{Recall} = \frac{TP}{TP + FN}, \qquad \text{FPR} = \frac{FP}{FP + TN}, \tag{17}$$

where TP = correctly detected inconsistent samples, FN = missed inconsistent samples, FP = consistent samples incorrectly predicted as inconsistent, and TN = correctly identified consistent samples. Each model is plotted as a single point; the ideal position is the upper-left corner (high recall, low FPR).

Confusion Matrices (Figure 5a). For each model, we construct an 8×8 confusion matrix $\mathbf{C}$ on segment-level classification, where $C_{ij} = |\{k : c_k = i \wedge \hat{c}_k = j\}|$ counts the number of samples

with ground-truth category i predicted as category j. Only correctly detected inconsistent samples (TP) are included. Each row is normalized to sum to 1 for visualization: $\bar{C}_{ij} = C_{ij} / \sum_{j'} C_{ij'}$.

Grounding Quantity vs. Quality (Figure 5c). For full-video temporal grounding, let N_i^{gt} and N_i^{pred} denote the number of ground-truth and predicted intervals for video i, respectively. The x-axis measures the average count difference:

$$\Delta N = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} (N_i^{\text{pred}} - N_i^{\text{gt}}), \tag{18}$$

where $\mathcal{V}$ is the set of videos with at least one predicted interval. Positive values indicate over-prediction; negative values indicate under-prediction. The y-axis measures proportion-based precision at IoU threshold α:

$$\text{Prec}(\alpha) = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \frac{|\mathcal{H}_i(\alpha)|}{N_i^{\text{pred}}}, \tag{19}$$

$$\mathcal{H}_i(\alpha) = \{ j : \max_k \text{IoU}(\hat{g}_j^{(i)}, g_k^{(i)}) \geq \alpha \},$$

where $\hat{g}_j^{(i)}$ is the j-th predicted interval and $g_k^{(i)}$ is the k-th ground-truth interval for video i. The ideal position is $\Delta N = 0$ (exact count match) with high precision.

Model	Cla.		Reasoning					
			B		R		M	
	R_d	P_d	R_d	P_d	R_d	P_d	R_d	P_d
Arc-Hunyuan-Video	9.0	5.0	1.13	0.63	13.2	7.4	13.7	7.7
OLA	1.1	3.5	0.18	0.57	2.2	6.8	1.7	5.2
SALMONN 2 -7B	0.0	12.5	0.00	0.38	0.0	5.2	0.0	5.2
SALMONN 2 -72B	1.4	29.5	0.07	1.48	0.7	15.0	0.7	15.6
Qwen3-Omni	16.4	50.0	0.66	2.00	6.5	19.7	7.1	21.6
MiMo-V2-Omni	19.8	37.5	0.96	1.82	8.3	15.8	8.4	16.0
Gemini 2.5 Pro	41.8	37.3	1.66	1.48	15.3	13.6	16.4	14.6
Gemini 3.1 Pro	37.3	40.8	1.63	1.79	13.9	15.2	14.8	16.2
AVID-Qwen	**33.1**	**36.5**	**3.69**	**4.07**	**15.8**	**17.4**	**15.4**	**17.0**

Table 13: Segment-level classification and reasoning metrics with recall- and precision-denominator variants. R_d: recall-denominator ($\div$(TP+FN)); P_d: precision-denominator ($\div$(TP+FP)).

C.6 Supplementary Results

Recall- and Precision-Denominator Metric Variants. The main results (Table 2) condition classification and reasoning metrics on correctly detected samples (TP). We additionally report two denominator variants: **recall-denominator** ($\div$ (TP+FN)), which normalizes by all actual inconsistent samples, and **precision-denominator** ($\div$ (TP+FP)), which normalizes by all predicted-positive samples. The conversion is:

$$M_{\text{Rec}} = M_{\text{TP}} \times \text{Recall}_{\text{det}}, \tag{20}$$

$$M_{\text{Pre}} = M_{\text{TP}} \times \text{Precision}_{\text{det}}. \tag{21}$$

Table 13 reports segment-level variants and Table 11 reports full-video variants.

Sentence-Level Semantic Similarity. To complement lexical metrics, we report Sentence Cosine Similarity using `all-MiniLM-L6-v2` (22M parameters, 384-dimensional embeddings) between model-generated and ground-truth reasoning. Sentence Cosine is computed on correctly detected samples (TP); for full-video tasks, unmatched GT intervals are assigned a score of 0. Results are shown in Table 14.

Model	Segment	Full-Video
Arc-Hunyuan-Video	39.0±12.4	5.3
OLA	35.6±13.4	7.7
SALMONN 2 -72B	43.7±14.1	-
Gemma4 E4B	53.8±12.5	-
Qwen3-Omni	53.7±12.3	8.3
MiMo-V2-Omni	48.5±14.7	16.0
Gemini 2.5 Pro	50.4±13.2	16.7
Gemini 3.1 Pro	52.4±12.6	18.2
AVID-Qwen	**56.7±16.8**	**28.3**

Table 14: Sentence Cosine Similarity ($\times 100$) on segment-level and full-video reasoning tasks.

Score	Annotation Instruction
5	The model reasoning is semantically consistent with the ground truth. All key information (inconsistency type, audio-visual evidence, and conclusion) is fully covered.
4	The model reasoning is largely correct and identifies the same inconsistency. Minor information is missing (e.g., one piece of evidence omitted) or contains slight redundancy, but the core judgment is accurate.
3	The model reasoning is partially correct. It captures some key points (e.g., correct inconsistency type but wrong evidence, or correct evidence but wrong category), with clear omissions or errors in the explanation.
2	The model reasoning is mostly incorrect. Only marginally related information is present (e.g., mentions audio or visual content but fails to identify the actual inconsistency or provides an unrelated explanation).
1	The model reasoning is completely incorrect, unrelated to, or directly contradicts the ground-truth reasoning.

Table 15: Human evaluation scoring rubric. Annotators rate each model-generated reasoning against the ground-truth on a 1–5 scale based on semantic coverage.

Category-wise Reasoning Quality. Table 12 reports reasoning quality (BLEU-4 / ROUGE-L / METEOR) broken down by the 8 inconsistency categories on the segment-level task, computed on correctly detected samples (TP) within each category.

D Additional Validity, Utility, and Diagnostic Results

D.1 General-Mixture Training Details

To evaluate AVID as a complementary sub-task rather than a stand-alone specialization corpus, we use Qwen2.5-Omni-3B and a 390,293-sample general mixture: LLaVA-Video (200,000), OLA-LLaVA-Video (120,925), AVQA (37,441), and FortisAVQA (31,927). We add 20,000 AVID segment samples without removing general examples. Training uses 32 A100 GPUs for approximately 14 hours. As reported in Table 4, the general mixture alone improves DailyOmni/WorldSense but leaves AVID detection at 51.5%; adding AVID raises it to 57.0%. DailyOmni decreases slightly relative to general-mixture-only training (63.83 to 63.17) but remains above the 62.32 base, while WorldSense improves from 47.19 to 47.52. We therefore interpret the result as preservation of general capability relative to the base model, not uniform improvement over the general mixture on every benchmark.

D.2 External Transfer to DAVE

We evaluate Qwen3-Omni and AVID-Qwen on the external DAVE benchmark (Radevski et al., 2025) using DAVE's evaluation protocol and no DAVE-specific tuning (Table 5). DAVE differs in source data and multiple-choice output format, making it a useful test of whether AVID training transfers beyond AVID-specific generation patterns and response formatting.

AVID-Qwen improves overall accuracy by 10.6 points, with particularly large gains on sound absence (+20.9) and sound discrimination (+19.6). This supports transferable audio-visual comparison capability, although we retain the in-domain AVID evaluation as the primary result.

D.3 Category–Reasoning Alignment Diagnostic

The original classification protocol uses one example per category and exact category-name matching. Because some labels have overlapping language (notably the two background-related categories), a semantically correct explanation can still map to the wrong string label. We diagnose this with two alternative protocols: three in-context examples per category, and rationale-only remapping by Gemini 3.1 Pro using five examples per category. The latter judge sees only the generated rationale, not the model's predicted category.

Protocol	Qwen3-Omni	AVID-Qwen
Name matching (1 example/category)	55.0	55.5
3 examples/category	50.5	58.5
Rationale remapping (5 examples/category)	36.4	62.5

Table 16: Category–reasoning diagnostic (segment classification accuracy, %).

Under rationale remapping, the base-to-AVID-Qwen gap widens from 0.5 to 26.1 points, indicating that AVID-Qwen's rationales frequently contain the correct category semantics even when name matching fails. Because this analysis changes the evaluation protocol and uses a proprietary judge, we treat it as a diagnostic rather than a replacement for the primary classification score.

Model	Low-level	High-level
Qwen3-Omni	1.7	65.6
Gemini 2.5 Pro	40.6	61.7
Gemini 3.1 Pro	53.6	59.5
AVID-Qwen	13.9	**75.4**

Table 17: Segment classification accuracy by inconsistency level, conditioned on correct detection (%).

D.4 Low-Level versus High-Level Categories

We further separate low-level physical synchronization categories (lip sync and temporal shift) from the remaining higher-level identity, semantic, emotional, spatial, and contextual conflicts. Table 17 reports classification accuracy over correctly detected inconsistent segments; these values are therefore conditional rather than end-to-end accuracy.

The aggregate score conceals a substantial capability split. AVID-Qwen's largest advantage lies in high-level semantic/contextual conflicts, whereas Gemini 3.1 Pro remains much stronger on low-level synchronization. This motivates reporting both groups alongside the aggregate metric.

E Human Evaluation Protocol

To assess the reliability of the automatic reasoning metrics, we compare BLEU-4, ROUGE-L, and METEOR with independent human ratings of model-generated explanations.

Two graduate students in computer science from the same research group, with research experience in multimodal learning, serve as annotators for the human evaluation (distinct from the 3 annotators used for benchmark construction in §A.4). Prior to the formal annotation, they completed a calibration session on 20 pilot cases (not included in the final evaluation) to align their understanding of the scoring rubric. We select 4 representative models spanning different capability levels: Gemini 3.1 Pro, Qwen3-Omni, OLA, and AVID-Qwen. For each model, we randomly sample 50 segment-level

cases from correctly detected inconsistent samples (True Positives), yielding 200 cases in total. All cases are shuffled and anonymized so that annotators cannot identify which model produced each output. Each case presents the ground-truth reasoning and the model-generated reasoning side by side; annotators are not shown the original video, category label, or model identity, and independently assign a score on a 1–5 scale following the rubric in Table 15. The scoring criteria evaluate semantic coverage: whether the model identifies the same inconsistency type, describes the same audio-visual evidence, and arrives at a consistent conclusion.

The two annotators achieve strong inter-annotator agreement: weighted Cohen's Kappa = 0.904, Spearman $\rho = 0.877$ ($p < 10^{-64}$), and ± 1 agreement = 98.0%. For each case, the final human score is the arithmetic mean of the two scores, rounded to the nearest integer. To produce Table 18, we group all 200 cases by their final human score (1–5) and compute the mean and standard deviation of each automatic metric (BLEU-4, ROUGE-L, METEOR) within each score group. The monotonic increase from score 1 to score 5 supports the correspondence between automatic metrics and human judgments of agreement with the reference explanations.

Score↑	**Count**	**BLEU-4↑**	**ROUGE-L↑**	**METEOR↑**
5	27	4.41±5.95	25.38±6.40	27.23±9.52
4	47	4.03±5.63	22.91±6.79	24.31±9.08
3	16	2.84±3.36	19.80±5.65	19.62±7.04
2	64	1.97±1.64	19.02±5.23	17.84±5.10
1	46	1.59±1.50	16.57±5.60	15.07±5.86

Table 18: Alignment between human judgments and automatic reasoning metrics.