AVID: Benchmarking Audio-Visual Inconsistency Understanding for Omni-Modal Language Models

Anonymous ACL submission

Abstract

Omni-modal large language models excel at temporally aligned tasks such as captioning and question answering, yet their ability to perceive audio-visual inconsistencies—a fundamental human cognitive capability critical for trustworthy multimodal understanding—remains largely unevaluated. To bridge this gap, we present AVID, the first large-scale benchmark for audio-visual inconsistency understanding in videos. AVID comprises 11.2K long-form videos (avg. 235.5 s) with 39.4K annotated inconsistency events and 78.7K segment clips, supporting evaluation across detection, temporal grounding, classification, and reasoning over 8 fine-grained inconsistency categories. 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. Extensive evaluation of state-of-the-art omni-models reveals substantial limitations in audio-visual inconsistency understanding and temporal localization. Our fine-tuned baseline, AVID-Qwen, achieves comprehensive improvements over the base model and surpasses the strongest closed-source omni-model, Gemini 3.1 Pro, in both temporal grounding (mIoU: 36.1% vs. 26.2%) and holistic understanding (SODA-m: 7.47 vs. 5.83), validating AVID as an effective testbed for advancing trustworthy omni-modal language understanding.

1 Introduction

The emergence of omni-modal large language models, such as Gemini 3.1 Pro (Gemini Team, Google DeepMind, 2025), MiMo-V2-Omni (Xiaomi AI Lab, 2025), and Qwen3-Omni (Xu et al., 2025), has significantly advanced machine perception. By jointly modeling visual, auditory, and textual signals, these models achieve strong performance on video captioning, visual question answering (VQA), and temporal action localization (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 immediately notice when an elderly man speaks with the voice of a young woman, or when audio narration conflicts with on-screen visual content, as illustrated in Figure 1.

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) focus on temporally aligned positive events where unimodal cues often suffice, while deepfake datasets (Yang et al., 2023; Cai et al., 2022, 2025) target artifact-level manipulations in short clips that can be detected without genuine cross-modal understanding (Xie et al., 2026; Haliassos et al., 2021). Neither setting adequately evaluates semantic-level audio-visual inconsistency perception in long-form videos (Table 1).

To address these limitations, we introduce **AVID**, the first benchmark for audio-visual inconsistency understanding in videos. 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), supporting evaluation across detection, temporal grounding, classification, and reasoning over 8 fine-grained inconsistency categories (Table 1).

We evaluate representative open-source and closed-source omni-modal models on AVID, revealing that even the best closed-source model

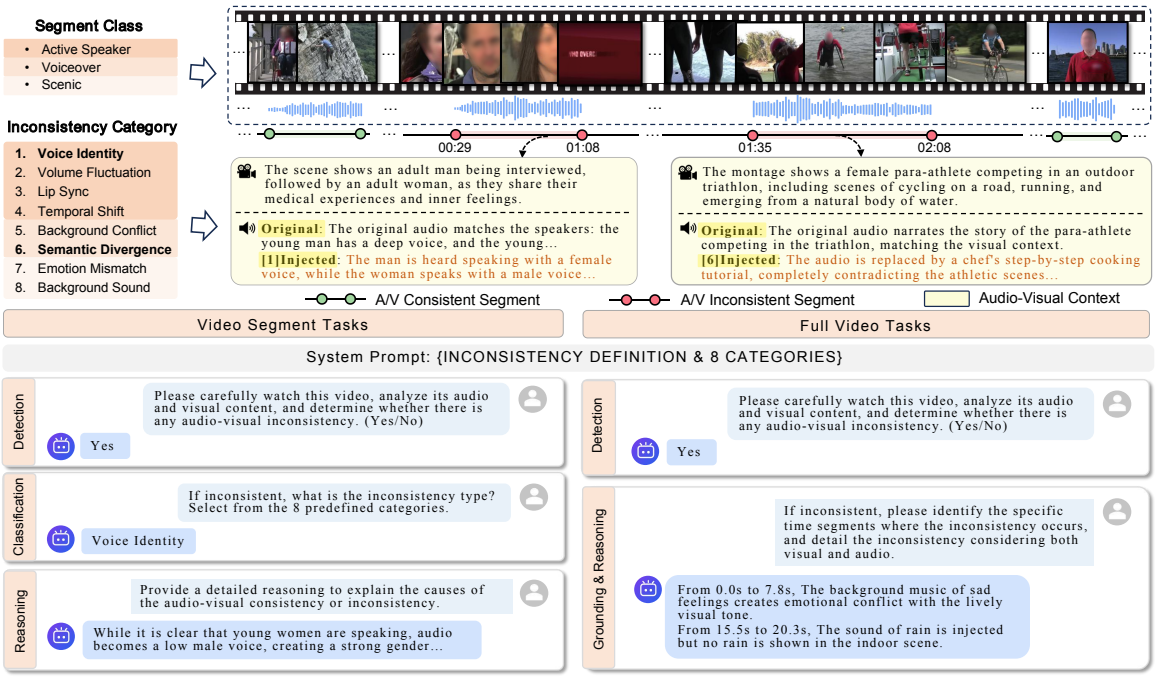


Figure 1: Overview of AVID, the first benchmark for audio-visual inconsistency understanding in videos, supporting both segment-level and full-video evaluation.

achieves only 69.7% segment detection accuracy and 26.2% full-video mIoU. We further introduce **AVID-Qwen**, a Qwen3-Omni model fine-tuned on AVID via a two-stage strategy, that surpasses the strongest closed-source model in temporal grounding and dense understanding. 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, the first benchmark for audio-visual inconsistency understanding, containing 11.2K full videos, 39.4K inconsistency event segments, and 78.7K clips with fine-grained category labels, temporal annotations, and natural-language reasoning.
- We evaluate leading omni-modal models on AVID and demonstrate that fine-tuning Qwen3-Omni (AVID-Qwen) surpasses the strongest closed-source model, Gemini 3.1 Pro, in temporal grounding (mIoU: 36.1% vs. 26.2%) and dense understanding (SODA-m: 7.47 vs. 5.83), validating AVID as effective supervision for learning fine-grained audio-visual inconsistency understanding.

2 Related Work

Audio-Visual Benchmarking. Early benchmarks such as AVE (Tian et al., 2018) addressed audio-visual event localization in untrimmed videos, while UnAV-100 (Geng et al., 2023) and Long-VALE (Geng et al., 2025) extended this line of work to dense multi-event localization in longer videos. However, these datasets focus exclusively on *aligned* events, assuming congruence between modalities, and thus cannot evaluate robustness to inconsistent or tampered content.

Deepfake and Inconsistency Detection. Deepfake datasets such as FakeAVCeleb (Yang et al., 2023) and LAV-DF (Cai et al., 2022) target facial manipulations and temporal forgery localization but rely on detecting unimodal artifacts (Aleem et al., 2026) rather than understanding cross-modal semantics. Cross-modal approaches such as AVFF (Oorloff et al., 2024) and AVoiD-DF (Yang et al., 2023) detect artifact-level inconsistencies (e.g., lip-sync errors, blending defects) that are identifiable without genuine multimodal reasoning, and lack natural-language reasoning annotations. In contrast, AVID targets *semantic-level* inconsistencies—such as voice-identity conflicts or scene-emotion mismatches—that require high-level cross-modal understanding, accompanied by fine-grained temporal annotations in long-form videos.

Dataset	Anno.	Video Form	#Videos	Avg. Len	Caption	A/V Inc.	#Inc. Cat.	Inc. Reasoning	Timestamp
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 Inconsistency Benchmark (Ours)									
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 the benchmark targets audio-visual inconsistency. #Inc. Cat.: number of inconsistency categories. V: visual, A: audio, S: speech. "–" denotes not applicable.

3 The AVID Benchmark

To construct AVID, we propose a scalable pipeline consisting of three stages: (i) data sourcing and class-aware segmentation (§3.1); (ii) strategy planning (§3.2); and (iii) inconsistency injection (§3.3). 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, 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 (Table 2). **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 7). This class-aware design ensures that injected inconsistencies are semantically grounded in the segment context, enabling evaluation of genuine cross-modal reasoning.

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

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 percepti-

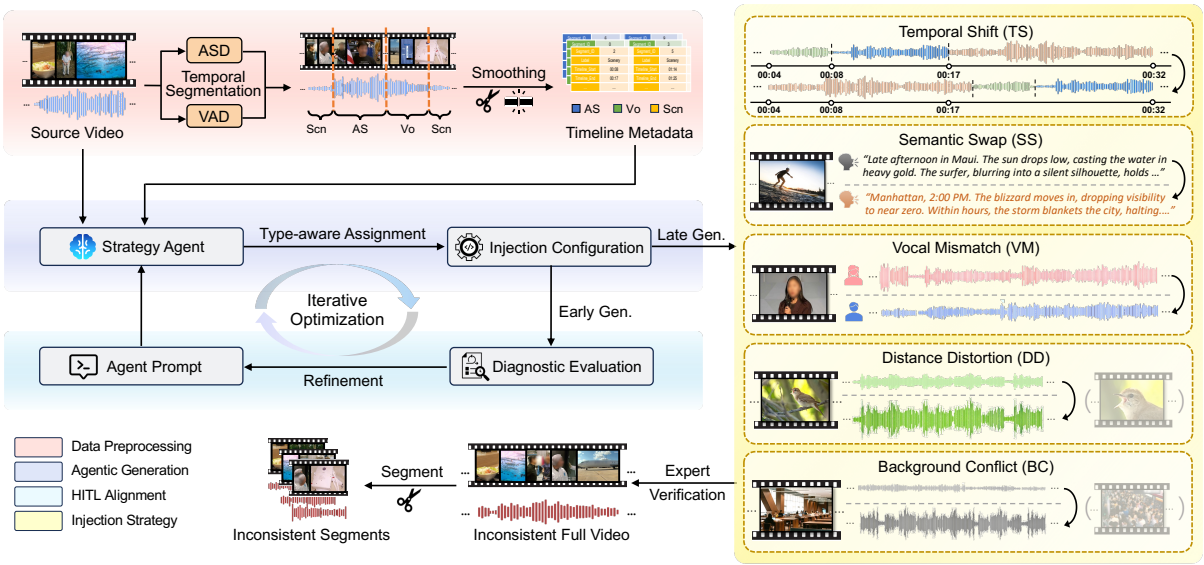


Figure 2: Overview of the AVID construction pipeline, consisting of three stages: (1) data preprocessing with temporal segmentation and class-aware labeling, (2) agentic generation with strategy planning and five specialized inconsistency injectors, and (3) human-in-the-loop alignment for quality control.

ble 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 Table 7), 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 (§3.4).

3.3 Inconsistency Injection

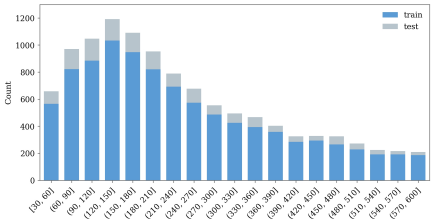
Given the injection plan produced by the Strategy Planning Agent, an Execution Agent parses each entry and routes the corresponding segment to one of five specialized injectors according to the specified inconsistency category (Figure 2, Stage 2): (1) **Temporal**—applies audio time-axis offset; (2) **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 7.

Crucially, all injectors operate exclusively on the audio stream; the original video frames, resolution, and duration are preserved throughout. This audio-only design is deliberate: it ensures that every injected inconsistency is inherently cross-modal and can only be detected by jointly reasoning over audio and visual signals, rather than from unimodal visual artifacts alone. Implementation details for each injector are provided in Appendix A.3.

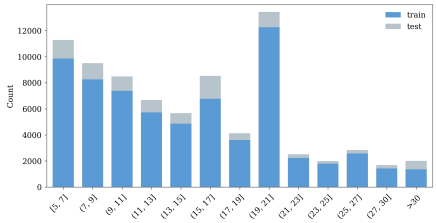
3.4 Human Effort and Subset Split

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 8). 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 $\kappa > 0.75$).

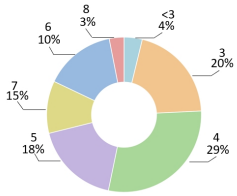
Through this pipeline, we construct 8,400 full videos with inconsistency events (39,361 inconsis-



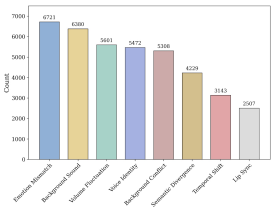
(a) Video duration distribution (train/test)



(b) Segment duration distribution (train/test)



(c) Inconsistency events per video



(d) 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.

tent 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.

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

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 3: Model comparison on AVID segment-level and full-video tasks. **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 ($\times 100$).

reasoning. The resulting AVID-Qwen serves as a transparent reference for comparing omni-models under a unified evaluation protocol.

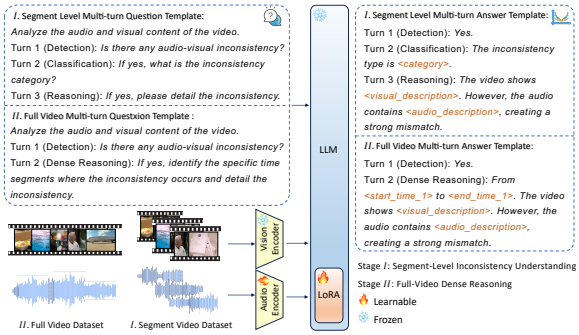


Figure 4: The AVID-Qwen architecture for omni-modal audio-visual inconsistency detection and reasoning.

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. All models are evaluated under a unified few-shot protocol with identical prompt templates. 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 metrics are therefore computed only on correctly detected samples (True Positives).

Metrics. We report **Accuracy** for detection and classification, **BLEU-4** (Papineni et al., 2002) / **ROUGE-L** (Lin, 2004) / **METEOR** (Banerjee and 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.

5.2 Main Results

Table 3 summarizes the results on both segment-level and full-video tasks.

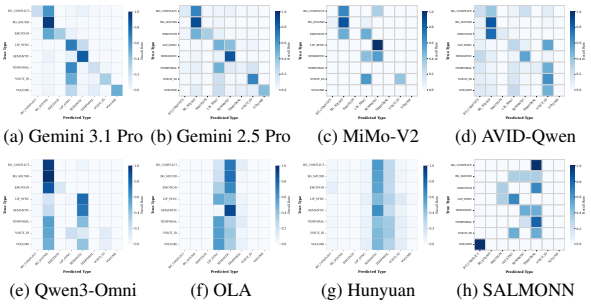
Overall Trends. Closed-source models (Gemini 2.5/3.1 Pro, MiMo-V2-Omni) consistently outperform open-source counterparts in detection (segment: 69.7% vs. 41–53%; full-video: 86.9% vs. 52–75%), while open-source models exhibit highly uneven capabilities—Qwen3-Omni shows balanced performance whereas OLA and ARC-Hunyuan-Video struggle across the board. Video-SALMONN 2 rarely predicts inconsistencies, artificially inflating its classification accuracy (Appendix C.4). AVID-Qwen demonstrates strong competitiveness with closed-source models, achieving 2.8× segment-level and 10.9× full-video BLEU-4 gains over the base model Qwen3-Omni,

while surpassing Gemini 3.1 Pro in both grounding (mIoU 36.1% vs. 26.2%) and dense understanding (SODA-m 7.47 vs. 5.83). These results confirm that AVID provides effective supervision for learning fine-grained audio-visual inconsistency understanding.

Segment-level Performance. For detection, the best closed-source model reaches 69.7% while open-source models cluster at 41–53%, indicating a persistent perception gap. Among correctly detected samples, classification narrows the closed-/open-source divide—Qwen3-Omni (55.0%) approaches Gemini (57.1%)—yet weaker models (OLA 7.9%, ARC-Hunyuan 11.1%) fail almost entirely, suggesting that even when these models can perceive inconsistencies, they lack the fine-grained audio-visual understanding needed to distinguish categories. Segment-level reasoning exhibits a similar pattern: Qwen3-Omni (METEOR 23.8%) approaches closed-source levels, but weaker models score below 18%, producing low-quality explanations. These results indicate that segment-level audio-visual understanding demands both perceptual sensitivity and fine-grained semantic reasoning, capabilities that most current models lack simultaneously.

Full-video Performance. Full-video evaluation amplifies the difficulty hierarchy. Detection remains accessible for strong models (Gemini 86.9%); note that Gemma4 E4B is excluded from full-video evaluation due to its 30-second audio input limit. Temporal grounding drops sharply—even the best zero-shot model achieves only 26.2% mIoU, and most open-source models fall below 10%. Full-video reasoning is universally weak: all zero-shot models score under 0.7 BLEU-4, confirming that generating faithful multi-event explanations over long videos remains an open challenge. The performance gradient—detection $\gg$ grounding $\gg$ reasoning—demonstrates that AVID effectively differentiates surface-level perception from deep temporal understanding. We further validate reasoning quality via sentence-level semantic similarity and category-wise breakdown (Appendix C.6), confirming consistent gains across all dimensions.

Reasoning Quality Validation. To validate the reliability of the automatic metrics, two annotators independently rate 200 randomly sampled reasoning outputs (spanning Gemini 3.1 Pro, Qwen3-Omni, OLA, and AVID-Qwen) on a 1–5 scale, achieving strong agreement (Cohen's $\kappa = 0.904$, Spearman $\rho = 0.877$; rubric in Appendix D). As shown in Table 4, higher human scores consistently correspond to higher automatic metrics, confirming that BLEU-4, ROUGE-L, and METEOR serve as reliable proxies for reasoning quality.



(a) Gemini 3.1 Pro (b) Gemini 2.5 Pro (c) MiMo-V2 (d) AVID-Qwen

(e) Qwen3-Omni (f) OLA (g) Hunyuan (h) SALMONN

Figure 5: Confusion matrices on segment-level 8-category inconsistency classification.

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

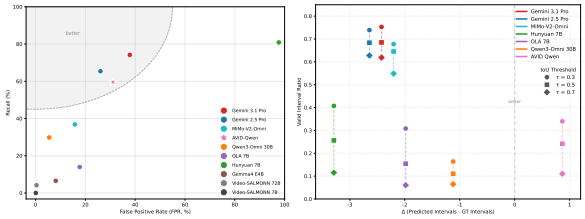
Table 4: Alignment between human scores and automatic metrics.

5.3 In-depth Analysis

Segment-level Task. Recall vs. False Positive Rate. As shown in Figure 6a, only the Gemini models initially occupy the ideal high-recall, low-FPR zone. After fine-tuning, AVID-Qwen enters this region, while other open-source models cluster toward extremes—either over- or under-predicting inconsistencies—revealing their difficulty in achieving balanced detection.

Fine-grained Classification. Figure 5 presents confusion matrices for 8-category classification. Gemini, MiMo-V2-Omni, and AVID-Qwen exhibit strong diagonal dominance, indicating balanced classification across all categories. In contrast, other open-source models distinguish only 1–2 categories (e.g., OLA concentrates on "lip sync" and "semantic"), revealing insufficient fine-grained discrimination.

Full-video Task. Figure 6b analyzes the trade-off between grounding quantity (difference between predicted and ground-truth interval counts) and precision (proportion of predictions exceeding an IoU threshold). Gemini and MiMo-V2-Omni adopt a conservative strategy, predicting fewer intervals but with high precision. Open-source models exhibit weaker and less stable behavior, achieving



(a) Recall vs. FPR (b) Grounding quantity vs. quality

Figure 6: (a) Segment-level recall vs. false positive rate (upper-left is ideal); (b) Full-video grounding quantity vs. quality (x-axis closer to zero and y-axis higher is ideal).

low precision despite predicting near ground-truth counts. AVID-Qwen displays a distinct pattern: it predicts more intervals than the ground truth, which lowers its proportion-based precision. However, this reflects a more comprehensive grounding strategy rather than weaker localization—AVID-Qwen achieves substantially higher recall than all other models (R@0.3: 55.2% vs. 30.9% for the next best; Table 3), indicating that it successfully identifies more valid inconsistency intervals while also attempting harder cases.

Model	Segment				Full-Video				
	Det.	Cla.	Reas.		Det.	Ground.		Reas.	Dense
	Acc	Acc	B	R	Acc	R@.5	mIoU	R	S_M
Qwen3-Omni	52.6	55.0	2.2	21.7	75.3	8.3	9.1	3.90	1.48
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

Table 5: Ablation study on training stages.

5.4 Ablation Study

We ablate the two-stage training strategy by comparing the full AVID-Qwen model with single-stage variants (Table 5). Training with Stage 1 alone yields strong segment performance but collapses full-video grounding (R@0.5: 4.6%); training with Stage 2 alone recovers full-video grounding but degrades segment classification (51.2%) and reasoning. The full two-stage model achieves the best balance, confirming that the two stages provide complementary supervision—segment-level training builds fine-grained local perception, while full-video training extends it to multi-event temporal reasoning. Notably, classification improves only marginally (55.0% $\rightarrow$ 55.5%) despite substantial gains in detection and reasoning, suggesting a misalignment between reasoning and category prediction that warrants future investigation.

Figure 7 presents representative outputs from Qwen3-Omni and AVID-Qwen on both tasks. The base model fails to detect segment-level inconsistencies and produces coarse full-video predictions with incorrect boundaries, whereas AVID-Qwen correctly identifies categories, generates precise reasoning, and localizes multiple events with fine-grained timestamps.

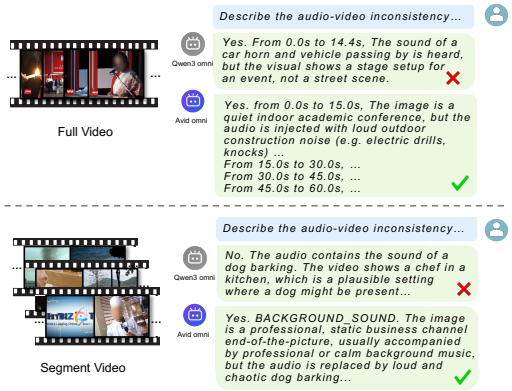


Figure 7: Qualitative comparison between Qwen3-Omni (base) and AVID-Qwen (fine-tuned) on segment-level and full-video tasks.

6 Conclusion

We present AVID, the first large-scale benchmark for audio-visual inconsistency understanding in long-form videos, comprising 11.2K videos with 39.4K annotated inconsistency events and 78.7K segment clips across 8 fine-grained categories. The benchmark is constructed through a scalable pipeline combining content-aware temporal segmentation with agentic inconsistency injection, and supports evaluation on detection, classification, temporal grounding, and reasoning at both segment and full-video levels. Extensive evaluation reveals that audio-visual inconsistency understanding remains an open challenge, with a clear capability hierarchy—detection $\gg$ grounding $\gg$ reasoning—where even the best closed-source model achieves only 69.7% segment detection accuracy and 26.2% full-video mIoU. Our two-stage fine-tuning strategy yields complementary gains, and the resulting AVID-Qwen surpasses Gemini 3.1 Pro in both temporal grounding (mIoU: 36.1% vs. 26.2%) and dense understanding (SODA-m: 7.47 vs. 5.83), validated by human evaluation. We additionally release standardized train/test splits in temporal video grounding and dense video captioning formats. Future directions include reasoning-category alignment objectives and leveraging AVID annotations to improve fine-grained temporal grounding.

Limitations

Audio-Only Manipulation. All inconsistency injections in AVID modify the audio stream while preserving the original video frames. This design does not cover video-side manipulations (e.g., face swapping, visual content insertion) or joint audio-visual manipulations, which constitute important real-world inconsistency scenarios.

Language and Cultural Coverage. All source videos and evaluation prompts are in English, which limits generalization to other languages and cultural contexts where audio-visual inconsistency cues may differ.

Category–Reasoning Misalignment. Classification accuracy improves only marginally (55.0% $\rightarrow$ 55.5%) despite substantial gains in detection and reasoning after fine-tuning, indicating that free-form reasoning and categorical prediction are not well aligned under the current training objective. We plan to first diagnose whether this misalignment stems from output formatting failures or genuine capability decoupling, by extracting implicit categories from generated reasoning texts and comparing them against model predictions. Based on this analysis, we will investigate targeted training strategies (e.g., joint reasoning-classification objectives, curriculum learning) to bridge this gap.

Single Fine-Tuning Backbone. AVID-Qwen is built exclusively on Qwen3-Omni. Whether the benefits of the two-stage training strategy generalize to other omni-modal architectures remains to be validated. We plan to extend the fine-tuning experiments to additional backbones in future work.

Ethics Statement

Source Data and Licensing. AVID is built on the publicly released LongVALE collection (Geng et al., 2025), whose source videos originate from YouTube. We release only derived annotations (temporal boundaries, segment classes, inconsistency labels, and natural-language reasoning) along with the audio manipulation pipeline; no raw video content is redistributed. Researchers reproducing AVID must comply with the YouTube Terms of Service and honor any platform-specific takedown requests.

Privacy and Identifiable Speakers. Source videos may contain identifiable individuals. The IdentityInjector applies low-level acoustic transforms (pitch, formant, tremolo) for category-level identity shifts (e.g., male$\rightarrow$female) and does not produce realistic impersonations of specific individuals. The SemanticInjector uses voice cloning only against the speaker reference within the local segment, never against external target identities. All manipulations are labeled within the benchmark as synthetic and are intended solely for evaluation.

Potential Misuse. The injectors could in principle be repurposed to fabricate misleading audio-visual content. We release AVID strictly for research purposes—benchmarking trustworthy omni-modal understanding, deepfake mitigation, and hallucination detection—and explicitly prohibit any application to non-consensual or deceptive media. The released package includes (i) documentation of intended use, (ii) per-event provenance metadata marking every manipulated segment as synthetic, and (iii) license terms prohibiting redistribution of generated audio outside research contexts.

Annotator Welfare and Consent. All annotation tasks—benchmark construction (Stage I and Stage II in §3.4), inter-annotator agreement studies (Appendix A.4), and human evaluation (Appendix D)—were performed by members of the same research group (graduate students and research assistants with backgrounds in computer science and multimodal learning), recruited through internal collaboration. Annotators were compensated at local fair-wage standards. All annotators provided informed consent to participate: they were briefed on the nature of the synthetic audio-visual content, the goals of the study, and how the resulting annotations would be used, and could withdraw or skip any sample at any time. Annotators reviewed only audio-visual content drawn from public web sources; no personally sensitive or harmful material was introduced by the pipeline.

Ethics Review. This study was determined to be exempt from formal institutional ethics review board (IRB) approval. The work involves annotation of publicly available media by informed, consenting research group members and does not constitute regulated human subjects research. No deceptive protocols, vulnerable populations, or personally sensitive data are involved.

Use of AI Assistants. Large language models (Gemini 3.1 Pro for strategy planning, Qwen3-TTS for speech synthesis) are used as components of

the AVID construction pipeline. AI assistants are not used to write the contents of this paper beyond standard grammar and clarity editing of author-written text.

References

Muhammad Aleem, Muhammad Umair, Muhammad Zubair, Rozeena Ibrahim, Muhammad Tahir Naseem, Muhammad Mohsin Raza, Muhammad Nadeem Ali, and Byung-Seo Kim. 2026. Seeing Through the Fake: Explainable AI With Multiple CNNs for Deepfake Detection. *IEEE Access*, 14:131–162.

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72.

Zhixi Cai, Kartik Kuckreja, Shreya Ghosh, Akanksha Chuchra, Muhammad Haris Khan, Usman Tariq, Tom Gedeon, and Abhinav Dhall. 2025. 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)*.

Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. 2022. 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)*, pages 1–10, Sydney, Australia. IEEE.

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

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. 2024. Panda-70M: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118.

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

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

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. 2025. ARC-Hunyuan-Video-7B: Structured video comprehension of real-world shorts. ArXiv:2507.20939.

Gemini Team, Google DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. ArXiv:2507.06261.

Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. 2023. 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)*.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. 2025. Long-VALE: 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)*.

Arushi Goel, Karan Sapra, Matthieu Le, Rafael Valle, Andrew Tao, and Bryan Catanzaro. 2024. OMCAT: Omni context aware transformer. ArXiv:2410.12109.

Google DeepMind. 2026. Gemma 4. Model card. Accessed: 2026-05-15.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2021. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, and 1 others. 2024. TrustLLM: Trustworthiness in large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

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

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

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

Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. 2021. 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)*.

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

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

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

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2025. Ola: Pushing the frontiers of omni-modal language model. ArXiv:2502.04328.

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

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

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

Silero Team. 2021. Silero VAD: Pre-trained enterprise-grade voice activity detector. GitHub repository.

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

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

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and 1 others. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

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. 2024. InternVid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations (ICLR)*.

Xiaomi AI Lab. 2025. MiMo: Unlocking the reasoning potential of language model – from pretraining to posttraining. ArXiv:2505.07608 [cs].

Shichuang Xie, Tong Qiao, Sheng Li, Xinpeng Zhang, Jiantao Zhou, and Guorui Feng. 2026. Deepfake detection in the AIGC era: A survey, benchmarks, and future perspectives. *Information Fusion*, 127:103740.

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, and 19 others. 2025. Qwen3-Omni Technical Report. ArXiv:2509.17765.

Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. 2023. AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake. *IEEE Transactions on Information Forensics and Security*, 18:2015–2029.

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

Hang Zhang, Xin Li, and Lidong Bing. 2023. 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)*, pages 543–553.

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 (≈1 s). **Step 3 (Classification):** Each time interval is labeled by combining VAD and face detection results: no speech → Class 3 (Scenic), speech + face → Class 1 (Active Speaker), speech + no face → 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.

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 8. 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.

A.3 Injector Details

Each injector follows a unified interface $\mathcal{I}_k : (V, A, M) \rightarrow (V', A', \Delta, \Theta_k)$, 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 6 and 7 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.

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 15–30 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 → 0.01) and "toward" (0.01 → 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 incon-

```
# 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 30 seconds long. If a segment exceeds 30 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 leads, negative = audio lags); `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; 15–30 s: 35–50; 20+ s: 50–70); `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 8: Prompt for strategy planning agent.

#	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 6: Overview of the five injection techniques in AVID.

Segment Class	Inconsistency Category	Injector	Description
Class 1 (Active Speaker)	VOICE_IDENTITY	IdentityInjector	Timbre/identity conflict (age/gender/voice-profile shift)
	VOLUME_FLUCTUATION	SpatialInjector	Loudness changes without corresponding visual movement
	LIP_SYNC	SemanticInjector	Mismatch between generated speech and lip motion
	TEMPORAL_SHIFT	TemporalInjector	Audio leads/lags video by 0.5–2.0 s
Class 2 (Voiceover)	BACKGROUND_CONFLICT	BackgroundInjector	Narration/music conflicts with visual scene context
	SEMANTIC_DIVERGENCE	SemanticInjector	Spoken content contradicts visual semantics
Class 3 (Scenic)	EMOTION_MISMATCH	BackgroundInjector	Background mood conflicts with scene affect
	BACKGROUND_SOUND	BackgroundInjector	Background sound conflicts with scene environment

Table 7: Class-conditional mapping between segment classes and applicable inconsistency categories.

sistency 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.

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. During fine-tuning, the vision encoder, audio encoder, and aligner are frozen; only the LLM backbone is updated 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 8 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 $\times$ 2 $\times$ 4)	
Training Epochs	2	10
Max Sequence Length	8192	
FPS Max Frames	12	
Video Max Pixels	50,176	
Hardware	6 $\times$ NVIDIA A100 80GB	
Distributed Training	DeepSpeed ZeRO-3	

Table 8: 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 9 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 us-

ing HuggingFace checkpoints. All models use temperature-based sampling with $T = 0.3$ for near-deterministic outputs. 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. All models receive a single-turn user message containing the evaluation prompt and video; prompt templates are provided in Figures 9 and 10. 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 a fixed decoding configuration with temperature set to 0.3 in the primary result table. Each sample is evaluated once (single run) as the low temperature setting ensures deterministic outputs with minimal variance. We evaluate on two complementary tasks: (i) segment-level detection, classification, and reasoning (Figure 9); and (ii) full-video detection with temporal grounding and dense reasoning (Figure 10). 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).

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.

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

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

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

Category Classification.

$$\text{Acc}_{\text{type}} = \frac{1}{N_{\text{inc}}} \sum_{i=1}^{N_{\text{inc}}} \mathbb{I}(\hat{c}_i = c_i). \tag{4}$$

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.

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

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

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), $\text{frag} = \#\text{chunks}/|m|$, $\gamma = 0.5$, $\theta = 3$.

Temporal Grounding.

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

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

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

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

Model	Provider/API	Version/Checkpoint	Context Len	Temp.	Top-p	Max Tokens	Frames/Rate	Audio Setting
Gemini 2.5 Pro	Google Vertex AI	gemini-2.5-pro-preview-0506	1M	0.3	-	4096	auto	Yes
Gemini 3.1 Pro	Google Vertex AI	gemini-3.1-pro-preview	1M	0.3	-	4096	auto	Yes
ARC-Hunyuan-Video	Local (HuggingFace)	TencentARC/ARC-Hunyuan-Video-7B	32K	0.7	0.9	2048	8fps	Yes
Qwen3-Omni	Local (HuggingFace)	Qwen3-Omni-30B-A3B-Instruct	8K	0.3	-	2048	12fps	Yes
OLA	Local	Ola-main/inference	8K	0.3	-	2048	12fps	Yes
Video-SALMONN 2-7B	Local	Video-SALMONN-2-7B	8K	0.3	-	2048	16kHz	Yes
Video-SALMONN 2-72B	Local	Video-SALMONN-2-72B	8K	0.3	-	2048	16kHz	Yes
MiMo-V2-Omni	OpenRouter	xiaomi/mimo-v2-omni	32K	0.3	-	4096	auto	Yes
Gemma4 E4B	Local (HuggingFace)	google/gemma-4-e4b-it	8K	0.3	-	2048	12fps	30s max
AVID-Qwen (w/ Stage 1)	Local (Qwen3-Omni)	Qwen3-Omni + LoRA Stage 1	8K	0.3	-	2048	12fps	Yes
AVID-Qwen	Local (Qwen3-Omni)	Qwen3-Omni + LoRA Stage 1+2	8K	0.3	-	2048	12fps	Yes

Table 9: Detailed test-time configuration for all evaluated large omni-models.

```
# Prompt for multi-turn audio-visual inconsistency judgment
```

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 0.5–2s (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 video shows a man speaking but the voice sounds like a woman).
- `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).

Multi-turn questions

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 9: Segment-level evaluation prompt for multi-turn audio-visual inconsistency judgment.

Model	Grounding								Reasoning						Dense	
	R@0.3		R@0.5		R@0.7		mIoU		B		R		M		S_M	
	R_d	P_d	R_d	P_d	R_d	P_d	R_d	P_d	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	0.01	0.01	1.06	1.06	0.50	0.50	0.39	0.39
OLA	8.9	11.3	4.4	5.6	1.7	2.2	5.5	6.9	0.11	0.13	2.27	2.87	1.21	1.53	0.75	0.95
SALMONN 2 -7B	0.0	25.0	0.0	25.0	0.0	0.0	0.0	13.8	0.00	0.16	0.01	6.45	0.00	4.95	0.00	0.38
Qwen3-Omni	12.3	9.3	8.3	6.3	4.9	3.7	9.1	6.9	0.25	0.19	3.89	2.94	2.48	1.87	1.48	1.11
MiMo-V2-Omni	12.2	23.3	11.6	22.2	9.9	18.9	10.2	19.5	0.28	0.53	2.87	5.50	2.65	5.08	2.80	5.37
Gemini 2.5 Pro	27.7	24.0	25.6	22.2	23.5	20.4	24.0	20.9	0.63	0.55	6.39	5.55	5.43	4.71	6.08	5.28
Gemini 3.1 Pro	29.5	25.6	26.8	23.3	24.2	21.0	25.0	21.7	0.55	0.48	6.71	5.83	4.88	4.24	5.56	4.83
AVID-Qwen	**53.5**	**43.5**	**38.0**	**30.9**	**18.5**	**15.1**	**35.0**	**28.4**	**2.65**	**2.15**	**14.74**	**11.98**	**14.55**	**11.82**	**7.25**	**5.89**

Table 10: Full-video grounding, reasoning, and dense understanding metrics with recall- and precision-denominator variants. R_d: recall-denominator (÷(TP+FN)); P_d: precision-denominator (÷(TP+FP)).

```
# Prompt for multi-turn audio-visual inconsistency detection in full videos
```

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 0.5–2s (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., video shows a man speaking but voice sounds like a woman).
- `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 background music of sad feelings creates emotional conflict with the lively visual tone
from 15.5s to 20.3s, The sound of rain is injected but no rain is shown in the indoor scene
```

Figure 10: Full-video evaluation prompt for multi-turn audio-visual inconsistency detection.

Model	1			2			3			4			5			6			7			8		
	B	R	M	B	R	M	B	R	M	B	R	M	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	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	1.1	14.4	11.8	1.0	15.0	11.9	1.2	16.4	11.9	1.2	16.8	12.5	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	0.5	9.6	9.5	1.5	16.4	20.9	0.8	13.3	15.7	2.9	20.1	18.3	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	-	-	-	-	-	-	1.7	12.7	13.9	2.1	21.0	22.0	1.8	23.6	19.4	-	-	-	-	-	-	-	-	-
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	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	1.5	17.5	17.6	2.4	22.2	22.5	1.1	19.6	17.3	1.8	19.9	20.3	2.9	24.8	24.6	-	-	-	4.5	25.8	24.2	-	-	-
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	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	2.2	21.6	23.0	2.6	22.2	22.9	1.7	19.1	22.0	2.4	19.7	20.8	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	**6.0**	**28.0**	**27.8**	**7.4**	**30.1**	**29.1**	**4.9**	**25.3**	**25.4**	**4.9**	**25.0**	**22.9**	**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 11: Category-wise segment-level reasoning quality (B/R/M = BLEU-4/ROUGE-L/METEOR). Categories: **1**=BG_CONFLICT, **2**=BG_SOUND, **3**=EMOTION, **4**=LIP_SYNC, **5**=SEMANTIC, **6**=TEMPORAL, **7**=VOICE_ID, **8**=VOLUME. Best per category in **bold**.

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:

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

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 §5.3.

Recall vs. False Positive Rate (Figure 6a). 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 5). 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 6b). 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:

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

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

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.

C.6 Supplementary Results

Recall- and Precision-Denominator Metric Variants. The main results (Table 3) 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:

$$M_{\text{Rec}} = M_{\text{TP}} \times \text{Recall}_{\text{det}}, \quad (20)$$

$$M_{\text{Pre}} = M_{\text{TP}} \times \text{Precision}_{\text{det}}. \quad (21)$$

Table 12 reports segment-level variants and Table 10 reports full-video variants.

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 12: 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)).

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 13.

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

Category-wise Reasoning Quality. Table 11 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.

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

D Human Evaluation Protocol

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 14. 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 ρ = 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 4, 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 confirms that automatic metrics are reliable proxies for reasoning quality.