

AVSCap: Orchestrating Audio-Visual Synergy for Omni-modal Video Captioning

Yanghai Wang^{1*}, Jiahao Wang^{1*}, Jiafu Tang^{1*},
 Yuanxing Zhang², Zhe Cao¹, Hanyan Bian¹, Zijie Zhang¹,
 Weiliang Luo¹, Zhiyu Pan¹, Zixuan Dong¹, Jiaheng Liu¹, Zhaoxiang Zhang^{3,†}

¹ NJU-LINK Team, Nanjing University ² Kling Team, Kuaishou Technology

³ Institute of Automation, Chinese Academy of Sciences

211300096@smail.nju.edu.cn

liujiaheng@nju.edu.cn

Abstract

Omni-modal video captioning is not merely combining visual captioning with audio transcription: a useful caption must describe how visual actions, speech, music, and sound effects co-evolve. Existing large multimodal models often fail at this relational step, treating audio and visual streams as loosely coupled observations, relying on automatic speech recognition, and under-specifying non-speech sounds and their links to visual events. We present **AVSCap**, a framework for audio-visual captioning centered on explicit cross-modal event binding. First, we construct **AVSCap-130K**, a tri-modal training corpus generated by a decoupled-then-fused pipeline that anchors visual and acoustic evidence before composing grounded omni-modal captions. Second, we train **AVSCap-7B**, a 7B captioner with a two-stage strategy: supervised fine-tuning establishes baseline capabilities, while sample-efficient reinforcement learning uses hybrid rewards to optimize acoustic completeness and audio-visual synergy. Our scaling analysis shows that reinforcement learning brings larger gains than increasing SFT data. Third, we introduce **AVSCapBench**, a benchmark that decomposes captions into visual, audio, and synergy events and evaluates them with fine-grained event recall. Experiments on AVSCapBench and external benchmarks show that AVSCap-7B improves non-speech audio coverage and cross-modal binding, delivering the best overall performance among evaluated open-source models.

1 Introduction

Large multimodal models (LMMs) have advanced video understanding toward omni-modal reasoning by jointly processing visual, audio, and textual signals (Xu et al., 2025a; Cheng et al., 2024; Li et al., 2025; Fu et al., 2025a; Liu et al., 2025). Among related tasks, omni-modal video captioning remains fundamental: an ideal caption must recognize visual events, speech, sound effects, and music, and describe how they co-evolve over time (Li et al., 2026; Chen et al., 2026).

Despite recent progress, current models suffer from an illusion of integration (Figure 1), with two bottlenecks. First, **modality isolation**: models process audio and visual streams as weakly coupled channels. They may mention a visual action and a co-occurring sound but fail to express their temporal relations (e.g., “while”, “as”), yielding information-rich yet relation-poor captions. Second, **speech-centric bias**: acoustic representations are dominated by automatic speech recognition (ASR), leaving non-speech sounds (e.g., collisions, ambient effects, music) under-specified.

Current benchmarks also struggle to evaluate these issues (Fu et al., 2025b; Li et al., 2024; Wu et al., 2025). Existing protocols often reward independent recognition: models can score highly by outputting visual and audio descriptions without demonstrating cross-modal synergy. Acoustic evaluation also mainly targets ASR correctness, largely ignoring background sound effects and music.

To address these gaps, we present **AVSCap**, a framework for explicit cross-modal event binding. First, we construct **AVSCap-130K**, a tri-modal corpus generated via a decoupled-then-fused pipeline that anchors unimodal evidence before composing audio-visual captions. Second, we train **AVSCap-7B** (based on Qwen2.5-Omni) by SFT and Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Guided by hybrid rewards, GRPO optimizes acoustic completeness and event binding. Finally, we introduce **AVSCapBench**, a human-curated benchmark decomposing captions into visual, audio, and synergistic events to evaluate audio sub-types and cross-modal synergy.

* Equal Contribution. † Corresponding Author.

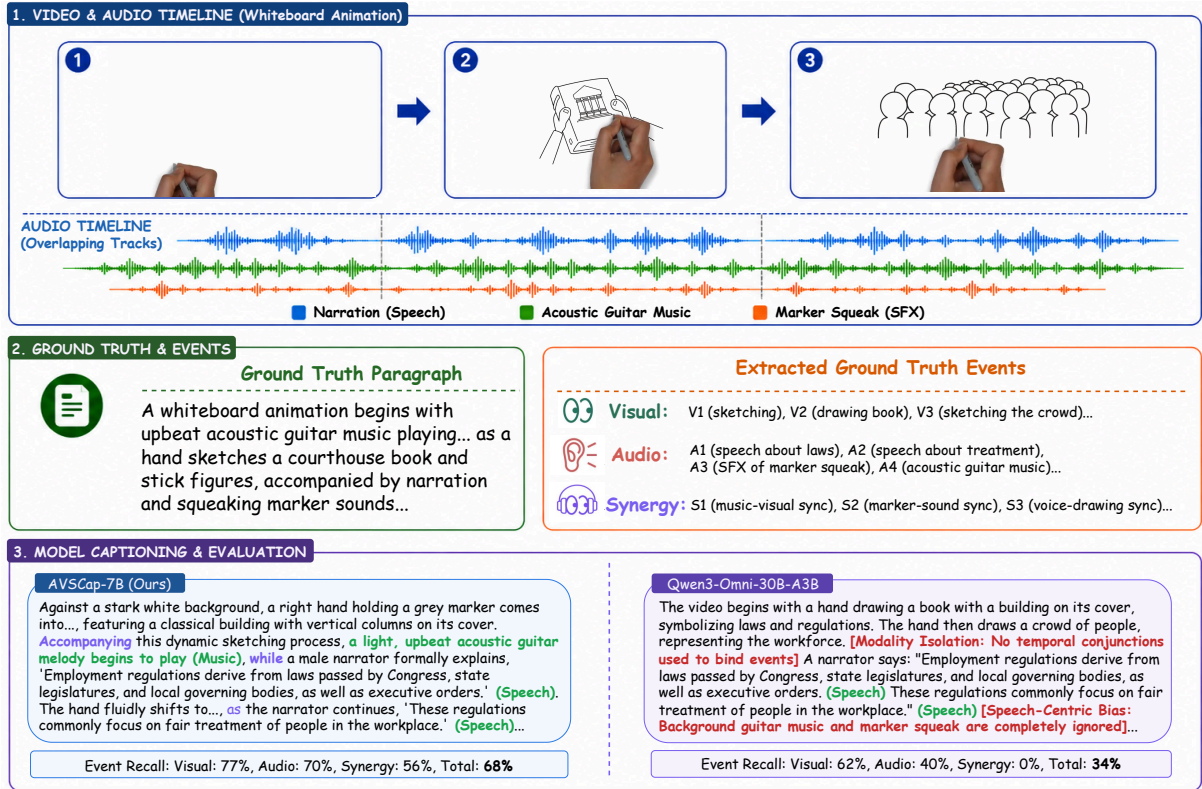


Figure 1: Overview of our evaluation protocol and core bottlenecks in omni-modal video captioning. In the bottom captions, color-coded text highlights key aspects: green denotes audio components, purple represents synergy conjunctions, and red highlights major model limitations.

Our main contributions are: **AVSCap-130K**. A tri-modal corpus of 130K orchestrated captions, providing explicit supervision for isolated perception and cross-modal grounding. **AVSCapBench**. A human-annotated benchmark (1,226 videos) featuring a fine-grained, event-based matching protocol. It explicitly evaluates visual, audio, and synergistic events, preventing models from achieving high scores via modality isolation. **AVSCap-7B**. A 7B captioner integrating SFT and GRPO to explicitly optimize acoustic completeness (including sound effects and music) and event binding.

2 Related Work

Audio-Visual Captioning. The emergence of omni-modal models (Comanici et al., 2025; Xu et al., 2025a; AI et al., 2025) has shifted video understanding from vision-centric perception to joint audio-visual modeling (Chen et al., 2026). Representative works include UGC-VideoCaptioner (Wu et al., 2025) and video-SALMONN-2 (Tang et al., 2025) for multimodal integration, AVoCaDO (Chen et al., 2026) for audiovisual temporal coherence, Omni-Captioner (Ma et al., 2026) and ASID-Captioner (Li et al., 2026) for detailed perception, TimeChat-Captioner (Yao et al., 2026) for structured multi-scene scripting, and OmniScript (Pu et al., 2026) for hierarchical script generation. While these methods improve audio-visual captioning, they do not center data construction, training, and evaluation around explicit event-level audio-visual binding. In contrast, AVSCap optimizes cross-modal synergy through a decoupled-then-fused data engine that anchors unimodal evidence before composing audio-visual captions.

RL for Video Captioning. RL (Schulman et al., 2017; Guo et al., 2025; Zheng et al., 2025; Gao et al., 2025) has become an important paradigm for aligning multimodal models with explicit objectives. CapRL (Xing et al., 2025) introduces verifiable rewards for caption generation, and VideoCap-R1 (Meng et al., 2025) uses GRPO to elicit structured thinking before captioning. AVoCaDO extends GRPO with content coverage and length regularization rewards, while TimeChat-Captioner trains its policy with task-specific rewards to align dense captions with fine-grained temporal boundaries. We adopt a regex-guided GRPO strategy with hybrid rewards to jointly optimize acoustic completeness and cross-modal event binding.

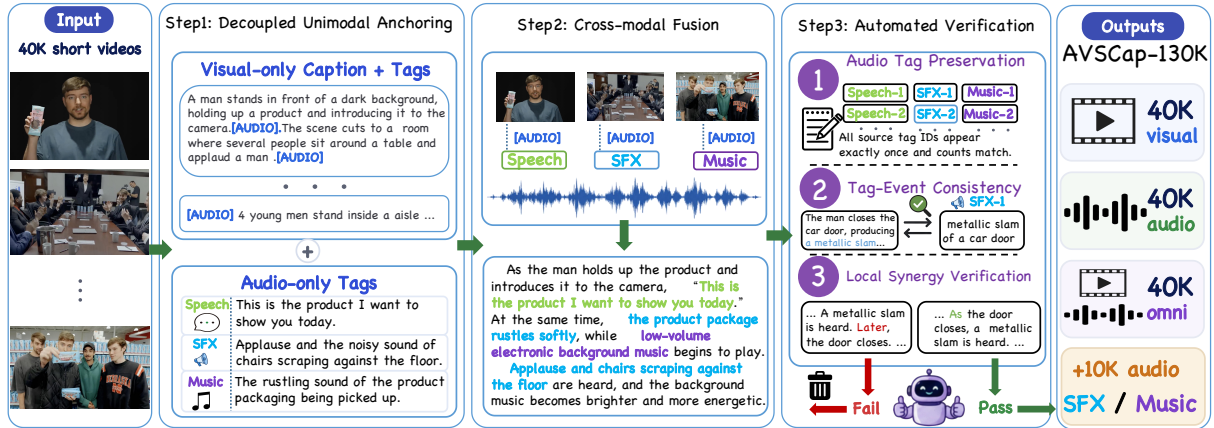


Figure 2: Overview of the AVSCap-130K data construction pipeline.

3 AVSCap

3.1 Data Construction Pipeline

3.1.1 Task Definition

Existing video captioning tasks mainly emphasize isolated visual or audio accuracy. We instead define a high-quality omni-modal caption by three criteria: (1) **Acoustic Completeness**, covering speech (Speech), sound effects (SFX), and music; (2) **Visual Completeness**, describing environments, characters, actions, object interactions, camera motion, and OCR; and (3) **Audio-Visual Synergy**, binding audio and visual events through coherent cross-modal relations, e.g., linking “a hammer falls” with “a striking sound (SFX)”.

3.1.2 Orchestrated Data Engine

To train AVSCap-7B, we construct AVSCap-130K, a 40K-video corpus with temporally grounded omni-modal captions. Videos are collected from AVoCaDO-107K (Chen et al., 2026), ASID-1M (Li et al., 2026), FineVideo (Farré et al., 2024), TimeChatCap-40K (Yao et al., 2026), and Movie101 (Yue et al., 2023), and filtered to clips shorter than 2.5 minutes. As shown in Figure 2, our decoupled-then-fused data engine has three stages: unimodal anchoring, cross-modal orchestration, and automated verification. Prompts used in this section are provided in Appendix D.2.

Step 1: Decoupled Unimodal Anchoring To reduce cross-modal interference during initial perception, we adopt a decoupled unimodal anchoring strategy using Gemini-3-Flash (Google DeepMind, 2026a). For the visual stream, the parser extracts scene-level attributes, including environments, characters, actions, object interactions, camera motion, and OCR text. When an audio cue is detected, the parser inserts an empty placeholder to indicate that the current visual segment has corresponding audio information, without describing the acoustic content itself. This preserves the visual-only nature of the caption while retaining temporal anchors for later cross-modal fusion.

For the audio stream, the parser separately extracts three acoustic categories: human speech (Speech), sound effects (SFX), and background music (Music). To standardize acoustic descriptions, SFX captions follow the annotation style of AudioCaps (Kim et al., 2019), while music captions follow the descriptive format of MusicCaps (Agostinelli et al., 2023).

Step 2: Cross-modal Fusion Given the unimodal anchors, we use Gemini-3-Flash to synthesize a coherent omni-modal caption. The fusion module aligns each visual placeholder with its corresponding audio description and inserts the audio content at the matched temporal position. To make cross-modal relations explicit, the fused caption uses temporal conjunctions such as “As”, “While”, “Simultaneously”, and “Accompanied by”. The original audio tags (Speech), (SFX), and (Music) are retained for later verification.

Step 3: Automated Verification To reduce hallucinations, duplication, and information loss during fusion, we apply a three-stage verification pipeline combining deterministic tag checks with local semantic validation.

Audio Tag Preservation. Since audio captions are chronologically ordered, we assign each audio event a unique tag ID before fusion, such as (Speech-1), (SFX-1), and (Music-1). After fusion, we verify that every source tag ID appears exactly once and that the retained tag count matches the number of source audio events for each type. Samples with missing, duplicated, or mismatched tags are discarded, as they indicate omitted, repeated, or misplaced audio events.

Tag-Event Consistency Check. Each retained audio tag must be attached to a local caption sentence and correspond to an integrated audio event. We use Qwen3-32B (Yang et al., 2025a) to verify whether the fused audio event remains semantically consistent with its source audio caption. Samples with unmatched tags or altered audio meanings are discarded to filter hallucinated, duplicated, or temporally misplaced audio insertions.

Local Synergy Verification. We further use Qwen3-32B to check whether each tag-anchored sentence contains both the audio event and its corresponding visual context, expressed through temporal or associative cues such as “while”, “as”, or “accompanied by”. Captions failing this check are removed, since they may preserve both modalities without forming valid cross-modal event binding.

3.2 Training AVSCap-7B

AVSCap-7B is built on Qwen2.5-Omni-7B and trained in two stages: SFT establishes baseline captioning, and GRPO improves acoustic completeness and audio-visual event binding. Additional details are provided in Appendix F.

3.2.1 Stage 1: Supervised Fine-Tuning (SFT)

We fine-tune Qwen2.5-Omni-7B with a multi-grained SFT dataset. Each video in **AVSCap-130K** provides three annotations: a visual caption, an audio caption, and a synergistic omni-modal caption. These prompt-response pairs enable both unimodal perception and cross-modal orchestration.

To improve non-speech audio understanding, we construct 10,000 audio-centric captions from AudioCaps (Kim et al., 2019) and MusicCaps (Agostinelli et al., 2023). Their raw audio tracks are processed with the same audio-stream pipeline in Section 3.1.2 to generate detailed (SFX) and (Music) captions. All video and audio captions are then jointly shuffled for unified SFT training. Ablations are provided in Appendix A.1.

3.2.2 Stage 2: GRPO Optimization

Although SFT provides strong captioning ability, the model still produces occasional repetitive outputs and weak audio-visual event binding. We further optimize **AVSCap-7B** with Group Relative Policy Optimization (GRPO) on 2,000 additional videos that do not overlap with the SFT training set. Their reference captions are constructed with the same data pipeline as **AVSCap-130K**. For each training instance, the policy samples a group of candidate captions, and optimization is guided by three rewards for length control, speech preservation, and audio-visual consistency.

Reward 1: Length Regularization (R_{len}). To prevent short outputs and repetition collapse, we assign full reward to captions within a valid length range and zero otherwise:

$$R_{\text{len}} = \mathbb{I}(\tau_{\min} \leq L_{\text{gen}} \leq \tau_{\max}), \quad (1)$$

where L_{gen} is the generated caption length. We set $\tau_{\min} = 200$ and $\tau_{\max} = 2048$ based on caption-length statistics and context-budget analysis, with details provided in Appendix B.

Reward 2: Regex-Anchored Speech Recall (R_{sp}). To encourage faithful preservation of speech content without expensive judge calls, we extract all (Speech) segments via regex matching and compare them with the reference speech segments. Before matching, punctuation is removed from both sides. For each aligned speech pair, we compute an LCS-based recall score:

$$R_{\text{sp}} = \frac{1}{M} \sum_{i=1}^M \frac{\text{LCS}(s_i^{\text{gen}}, s_i^{\text{ref}})}{|s_i^{\text{ref}}|}, \quad (2)$$

where M is the number of reference speech segments, s_i^{gen} and s_i^{ref} denote the generated and reference speech segments after tag-based alignment, and $\text{LCS}(\cdot)$ denotes the length of their longest common subsequence. This reward penalizes omitted, reordered, or substantially rewritten speech while remaining fully rule-based.

Reward 3: Cross-modal Synergy Recall (R_{syn}). To optimize the correctness and consistency of audio-visual event binding, we decompose the reference caption into synergy events and use GPT-5 (Singh

Benchmark	#Videos	Avg. D	A	V	AVS	Sub-T
UGC-VidCap	1,000	23.9s	✓	✓	✗	✗
Omni-Cloze	2,340	34.2s	✓	✓	✓	✗
v-SALMONN2	483	50.8s	✓	✓	✗	✗
AVSCapBench (Ours)	1,226	60.1s	✓	✓	✓	✓

Table 1: Comparison of audio-visual video captioning benchmarks. Avg. D: Average Duration; A: Audio; V: Visual; AVS: Audio-Visual Synergy; Sub-T: Audio Sub-types. “v-SALMONN2” is abbreviated for the video-SALMONN-2-testset. ✓: supported; ✗: not supported.

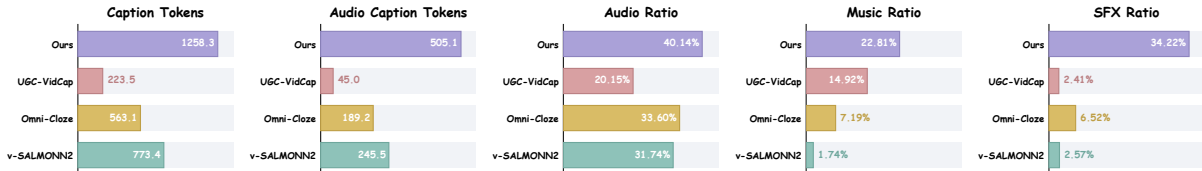


Figure 3: Comparison of caption characteristics across benchmarks. Metrics include average token lengths for total and audio-specific descriptions (Caption/Audio Tokens), the ratio of audio tokens (Audio Ratio), and internal distribution of music and sound effects (Music/SFX Ratio). v-SALMONN2 refers to video-SALMONN-2 testset.

et al., 2025) to judge whether each event is covered by the generated caption. The reward is computed as event-level recall:

$$R_{\text{syn}} = \frac{|\mathcal{E}_{\text{covered}}|}{|\mathcal{E}_{\text{syn}}|}, \quad (3)$$

where \mathcal{E}_{syn} is the set of reference synergy events and $\mathcal{E}_{\text{covered}}$ is the subset covered by the generated caption. This reward encourages preserving both modalities and their temporal correspondence.

Total Reward. The final reward is the sum of the three components:

$$R_{\text{total}} = R_{\text{len}} + R_{\text{sp}} + R_{\text{syn}}. \quad (4)$$

3.3 The AVSCapBench

To rigorously evaluate the capabilities of omni-modal models in video understanding and audio-visual synergy, we introduce the AVSCapBench.

3.3.1 Benchmark Construction

AVSCapBench consists of 1,226 manually annotated video clips collected from YouTube, TikTok, and Video-MME (Fu et al., 2025b), lasting 30 to 120 seconds. It covers diverse domains, including movies, vlogs, gaming, sports, and news. A complete annotation case is provided in Appendix C. Table 1 and Figure 3 compare AVSCapBench with existing benchmarks, and Figure 4 summarizes its statistics. We construct AVSCapBench through a three-stage human-in-the-loop pipeline.

Automated Segmentation. We first use Gemini-3-Flash (Google DeepMind, 2026a) to identify segmentation timestamps that preserve both visual coherence and audio continuity. The original videos are then divided into shorter clips for annotation.

Segment-level Human Annotation. Human annotators independently caption each clip, covering visual content like scenes, characters, actions, objects, and OCR text alongside speech, sound effects, and music. Annotators are instructed to place audio descriptions near their corresponding visual events and to ensure that each caption reflects audio-visual consistency throughout the segment.

Merging and Cross-Verification. Segment-level captions are sequentially merged into dense full-video captions. A second group of annotators then cross-checks them to identify omissions, hallucinations, and incorrect audio-visual alignments. To stress-test fine-grained acoustic coverage, we retain only samples that contain all three audio categories—Speech, SFX, and Music; samples that do not satisfy this requirement are discarded.

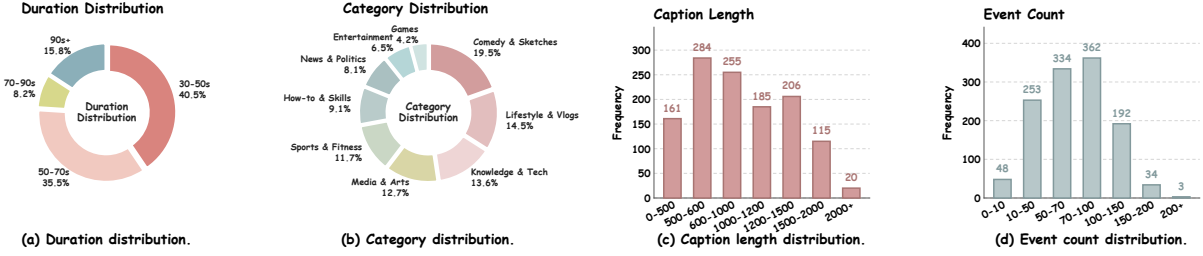


Figure 4: Statistics of AVSCapBench. (a) Videos span durations from 30 to 120 seconds. (b) Diverse categories. (c) Caption lengths vary across the dataset. (d) Distribution of atomic events per video.

3.3.2 Evaluation Protocol

Inspired by DREAM-1K (Wang et al., 2024), we adopt a fine-grained event-based matching protocol, as illustrated in Figure 1. We utilize GPT-5 (Singh et al., 2025) to decompose each ground-truth caption directly into a structured set of atomic events \mathcal{E} categorized across three distinct modality types. **Visual events** \mathcal{E}_{visual} describe objective visual content, including entities, actions, interactions, and scene states. **Audio events** \mathcal{E}_{audio} cover auditory information, including speech, sound effects, and background music, organized in chronological order. **Synergy events** $\mathcal{E}_{synergy}$ bind a visual event with its corresponding audio cue when the two are temporally aligned, thereby capturing cross-modal audio-visual relations. We empirically demonstrate the limitations of traditional n-gram overlap metrics and justify our transition to event-based recall in Appendix A.3.

During evaluation, we use Gemini-3.1-Pro (Google DeepMind, 2026b) as the judge model. For each event $e_i \in \mathcal{E}$, the judge determines whether the model-generated caption \hat{C} semantically covers the event. We then compute recall independently for each event type:

$$\text{Recall}_{\text{type}} = \frac{1}{|\mathcal{E}_{\text{type}}|} \sum_{e_i \in \mathcal{E}_{\text{type}}} h(e_i, \hat{C}), \quad (5)$$

where $h(e_i, \hat{C}) = 1$ if the judge determines that \hat{C} covers event e_i , and $h(e_i, \hat{C}) = 0$ otherwise. The event type is defined as $\text{type} \in \{\text{visual}, \text{audio}, \text{synergy}\}$. Since the numbers of events differ across modality types, the overall benchmark score is computed as the event-count-weighted average of the three recall metrics. The complete set of system instructions and evaluation prompts is detailed in Appendix D.

4 Experiments

4.1 Main Results

We evaluate 13 leading omni-modal models, including Gemini-3-Pro (Google DeepMind, 2026c), Gemini-3-Flash (Google DeepMind, 2026a), Qwen3-Omni (Xu et al., 2025b), Qwen2.5-Omni (Xu et al., 2025a), ARC-Hunyuan-Video (Ge et al., 2025), HumanOmniV2 (Yang et al., 2025b), MiniCPM-o (Yao et al., 2024), video-SALMONN-2 (Tang et al., 2025), ASID-Captioner (Li et al., 2026), AVoCaDO (Chen et al., 2026), and UGC-VideoCaptioner (Wu et al., 2025). We provide the detailed evaluation settings and an extended set of evaluation results in Appendix E.1 and E.2, respectively. The main results in Table 2 yield several key observations: (1) **Audio-visual synergy remains the weakest capability.** Synergy Recall is consistently the lowest dimension across all models. Open-source models often exhibit near-zero synergy despite strong unimodal perception, indicating a severe lack of temporal alignment. (2) **Specialized models outperform general omni-modal models.** Task-specific captioners (e.g., AVoCaDO) surpass general omni-modal models (e.g., Qwen2.5-Omni), underscoring the value of task-oriented training. (3) **Audio understanding is highly inconsistent** across open-source systems, with music and SFX remaining particularly challenging. (4) **AVSCap-7B achieves competitive performance**, outperforming open-source baselines and approaching commercial systems, validating the effectiveness of our curated data and GRPO-based optimization.

4.2 Cross-Benchmark Evaluation

To assess generalization, we evaluate AVSCap-7B on UGC-VideoCap (Wu et al., 2025), Daily-Omni (Zhou et al., 2025), and Omni-Cloze (Ma et al., 2026). Detailed descriptions of these external benchmarks can be found in Appendix G. As shown in Table 3, AVSCap-7B demonstrates robust transferability,

Model	Visual	Audio				Synergy	Total
		Speech	Music	SFX	Overall		
<i>Closed-Source Models</i>							
Gemini-3-Pro	60.43	79.81	39.52	27.77	71.29	48.88	60.97
Gemini-3-Flash	58.14	79.78	39.46	32.34	72.65	48.94	60.54
<i>Open-Source Models</i>							
AVoCaDO-7B	50.59	70.42	38.71	19.25	61.07	29.13	49.31
ASID-Captioner-7B	47.42	68.73	30.50	17.91	59.02	24.84	45.94
ASID-Captioner-3B	43.63	66.95	27.06	17.31	57.53	21.36	43.03
Qwen3-Omni-30B-A3B-Instruct	41.85	49.08	9.34	8.68	39.17	16.19	35.29
video-SALMONN-2-7B	39.05	46.76	13.76	8.71	36.52	12.43	32.02
UGC-VideoCaptioner-3B	33.24	21.30	22.00	11.48	20.77	10.43	24.24
Qwen2.5-Omni-7B	34.78	13.92	4.02	7.22	13.71	7.00	21.53
ARC-Hunyuan-Video-7B	20.68	16.49	3.93	1.97	11.41	4.52	14.49
HumanOmniV2-7B	27.78	4.60	1.58	2.46	4.41	2.42	14.10
MiniCPM-o-2.6-8B	24.61	6.75	3.31	3.92	6.13	3.78	13.66
AVSCap-7B (Ours)	59.33	69.45	40.36	30.82	64.30	57.70	60.44

Table 2: Main results on AVSCapBench. All values are Recall (%).

Model	UGC-VideoCap				Caption-to-QA
	Aud. ↑	Vis. ↑	Det. ↑	Avg. ↑	Daily-Omni ↑
<i>Closed-Source Commercial Models</i>					
Gemini-3-Pro	80.4	84.7	80.6	81.9	66.4
Gemini-2.5-Pro	69.5	74.7	73.7	72.6	60.2
<i>Open-Source Models</i>					
HumanOmniV2-7B	45.6	66.3	59.5	57.1	8.2
ARC-Hunyuan-Video-7B	52.7	56.0	55.8	54.8	8.6
MiniCPM-o-2.6-8B	38.6	68.5	57.7	54.9	9.8
Qwen2.5-Omni-7B	46.9	66.1	60.0	57.7	13.4
UGC-VideoCaptioner-3B	61.4	58.4	57.5	59.1	17.0
video-SALMONN-2-7B	61.8	71.4	68.5	67.2	29.9
Qwen3-Omni-30B-A3B	67.5	74.8	72.3	71.5	17.5
AVoCaDO-7B	73.0	74.6	71.8	73.2	50.1
ASID-Captioner-3B	78.6	84.8	80.2	81.2	55.5
ASID-Captioner-7B	79.1	84.4	80.2	81.2	61.2
AVSCap-7B (Ours)	82.9	81.1	83.2	82.4	66.6

Table 3: Model performance on the audiovisual captioning and QA benchmarks. For Daily-Omni, we report QA performance by Gemini-2.5-Pro based on captions.

outperforming all open-source baselines and approaching proprietary models on both UGC-VideoCap and Daily-Omni. Furthermore, Table 4 confirms its superiority on Omni-Cloze, particularly in the Audio-Visual subset, validating that our synergy-focused training successfully transfers to dense cross-modal reasoning tasks.

Model	Audio	Visual	Audio-Visual	Overall
ASID-Captioner-7B	46.3	42.3	47.7	44.4
AVoCaDO-7B	49.7	41.5	48.6	45.3
Qwen2.5-Omni-7B	10.4	12.9	18.9	12.9
AVSCap-7B (Ours)	45.6	52.1	57.8	50.8

Table 4: Detailed results on Omni-Cloze.

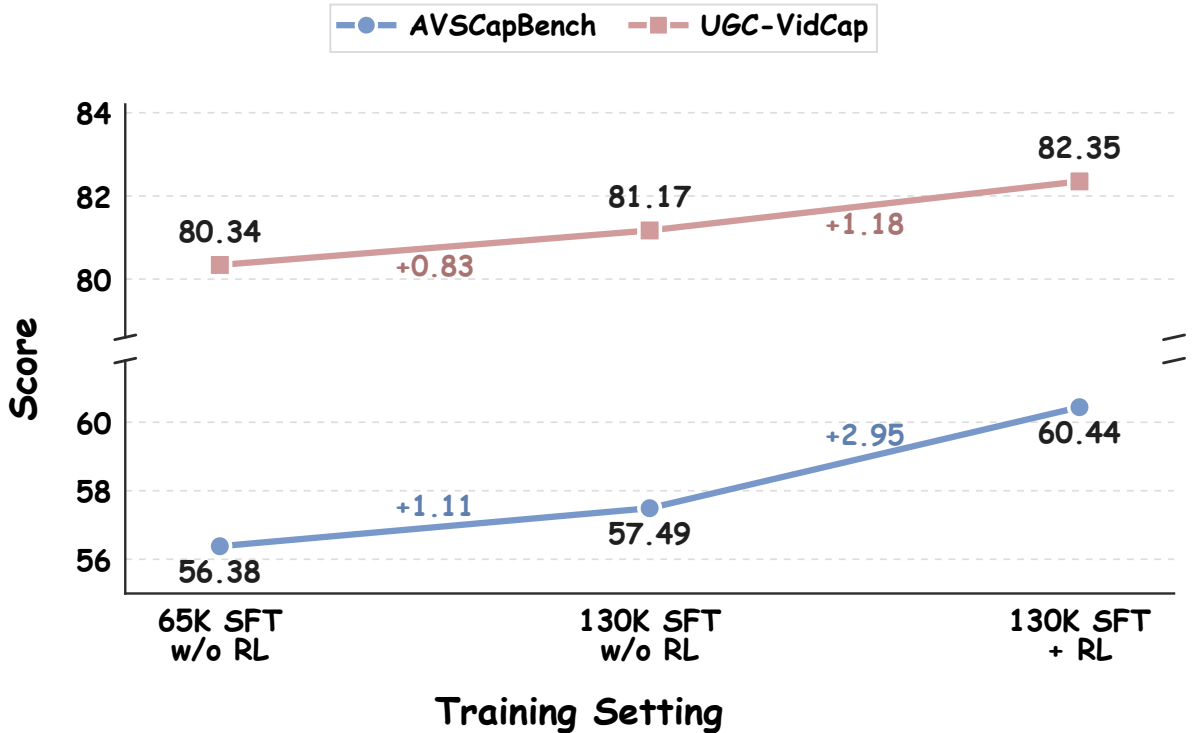


Figure 5: Ablation on SFT data scaling and GRPO optimization. The 65K SFT setting is randomly sampling 50% of each data from AVSCap-130K.

4.3 Ablation Studies

Efficacy of GRPO vs. SFT Data Scaling. To examine whether the gains come from reinforcement learning or simply from more supervised data, we compare GRPO with SFT data scaling in Figure 5. Starting from the full AVSCap-130K training set, we construct a half-scale SFT variant by randomly sampling 50% of each data type, resulting in a 65K-example subset. Scaling SFT from 65K to 130K yields only marginal improvements on AVSCapBench and UGC-VideoCap. In contrast, applying GRPO on top of the 130K SFT model with only 2K additional optimization videos brings substantially larger gains, indicating that direct policy optimization is more effective than pure SFT data scaling for improving audio-visual synergy.

Effect of Video Duration. We evaluate three open-source models across four duration intervals (30–50s, 50–70s, 70–90s, and over 90s) in Figure 6. All evaluated models show a consistent downward performance trend as video length increases. This universal decline highlights a key limitation of current open-source omni-modal architectures: they struggle with long-context processing, making it difficult to maintain robust and aligned audio-visual perception across extended temporal horizons.

Effect of Frame Sampling Rate. We analyze the impact of visual sampling rates (0.5, 1, 2, and 4 FPS) in Figure 7. Open-source models show a non-monotonic trend, peaking at 2 FPS but degrading at 4 FPS, which suggests that moderate frame rates optimize perception while overly dense sampling introduces context redundancy and distribution shifts. In contrast, Gemini-3-Flash continuously improves up to 4 FPS, demonstrating stronger long-context robustness and redundancy filtering.

Modality Shielding and Leakage. Beyond overall captioning quality, an omni-modal model should also support controllable modality-specific generation: when asked to describe only the visual or audio stream, it should avoid leaking information from the suppressed modality. This ability is important for controllable captioning and for verifying whether the model truly separates unimodal evidence before cross-modal fusion. To evaluate this, we sample 100 videos from AVSCapBench that contain no subtitles or on-screen dialogue cues, preventing models from inferring speech content from visible text. For each video, we prompt each model to generate either a visual-only or an audio-only caption. We then use Gemini-3.1-Pro to check whether the generated caption contains information from the suppressed modality. A sample is counted as leakage if cross-modal content is detected, and the Modality Leakage Rate is computed as the percentage of leaked samples over the 100 videos. As shown in Table 5, general and task-specific baselines frequently leak suppressed modality information, whereas

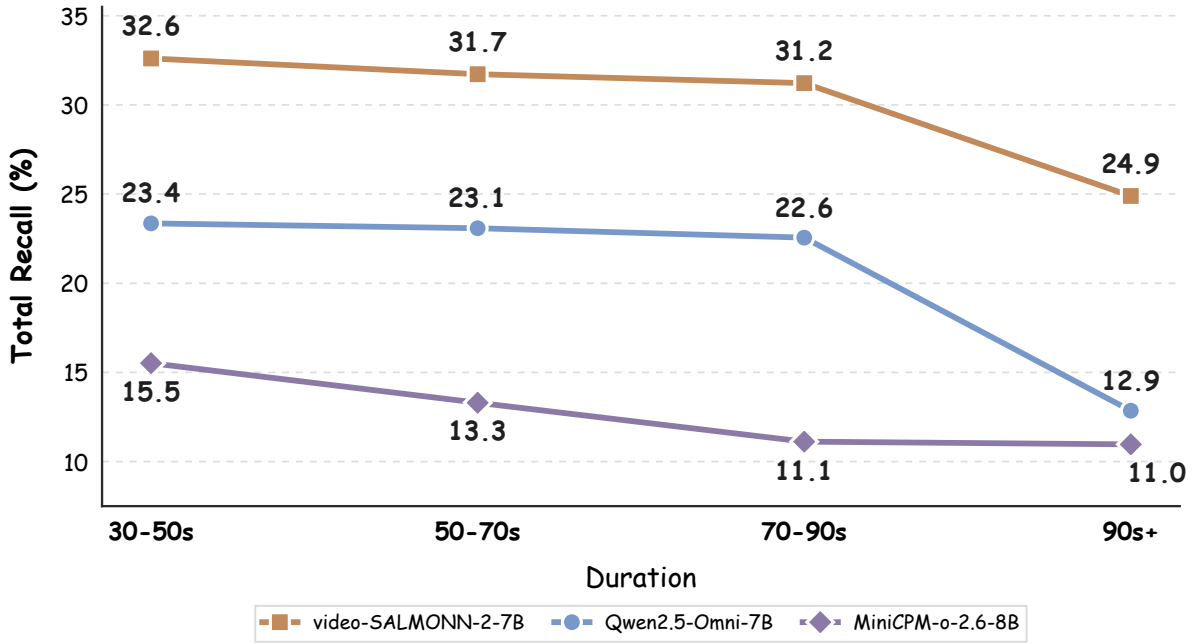


Figure 6: Effect of video duration on AVSCapBench.

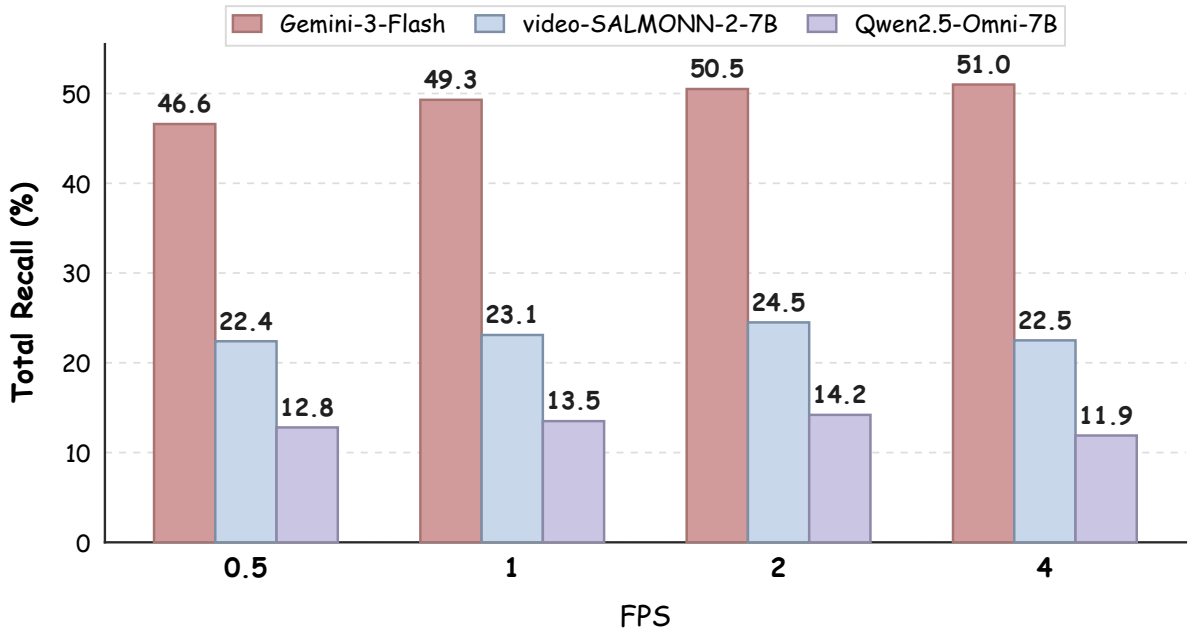


Figure 7: Effect of FPS on AVSCapBench.

AVSCap-7B achieves substantially lower leakage rates, demonstrating stronger modality isolation and instruction-controllable generation. We further analyze prompt sensitivity in Appendix A.2.

Judge-Human Agreement. To validate automated evaluation reliability, we assess the alignment between human annotations and three LLM judges (Gemini-3.1-Pro, DeepSeek-V4-Pro (DeepSeek, 2026), and Qwen3.5-27B (Qwen Team, 2026)) across visual, audio, and synergy dimensions on 200 benchmark samples (Table 6). All LLM judges demonstrate strong agreement with human ratings, with Gemini-3.1-Pro achieving the highest consistency; we thus adopt it as our primary automated evaluator. Importantly, we observe that while these judges exhibit different levels of agreement, they yield identical partial orderings for the evaluated models, ensuring the robustness of our ranking results (refer to Appendix E.3 for detailed statistics and agreement calculations).

Model	Audio Leakage ↓	Visual Leakage ↓
Gemini-3.1-Pro	38.0%	48.0%
ASID-Captioner-7B	49.0%	29.0%
Qwen2.5-Omni-7B	72.0%	25.0%
Qwen2.5-Omni-3B	73.0%	29.0%
AVoCaDO-7B	99.0%	89.0%
AVSCap-7B (Ours)	4.0%	7.0%

Table 5: Modality Leakage Rate on 100 videos.

Model	Visual	Audio	Synergy	Total
Gemini-3.1-Pro	0.97	0.95	0.91	0.94
DeepSeek-V4-Pro	0.94	0.89	0.87	0.90
Qwen3.5-27B	0.92	0.88	0.84	0.88

Table 6: Agreement between automated evaluation and human evaluation across different judges.

Error Analysis. To further analyze model failures, we randomly sample 200 cases and inspect the unmatched events across multiple models. We categorize errors according to the three event types used in our evaluation. Visual errors are divided into missing visual information and incorrect visual descriptions. Audio errors are divided into incorrect acoustic descriptions and partial audio omissions. Synergy errors are grouped into three types: missing audio-visual relations, incorrect cross-modal binding, and complete event omission. Overall, weaker open-source models often miss audio and synergy events entirely, while stronger models tend to fail through fine-grained incorrect descriptions or imperfect binding. Detailed statistics and examples are provided in Appendix H.

5 Conclusion

We present AVSCap, a unified framework for fine-grained audio-visual synergy that addresses modality isolation and speech-centric bias in omni-modal video captioning. We construct AVSCap-130K, a tri-modal training corpus enforcing isolated unimodal perception before cross-modal grounding, and train AVSCap-7B with a two-stage SFT-GRPO paradigm to optimize event binding. We further introduce AVSCapBench, a human-curated benchmark with a fine-grained, event-based matching protocol for evaluating visual, audio, and synergy dimensions. Experiments show that AVSCap-7B substantially outperforms open-source baselines and approaches commercial-grade performance.

Limitations

While AVSCap demonstrates promising performance, several limitations remain. First, model capability declines as video duration increases, reflecting the general difficulty open-source architectures face in modeling long-term spatiotemporal audio-visual dependencies. Second, our dataset and benchmark are currently restricted to English; extending the framework to multilingual settings, non-English dialogue, and broader cultural contexts remains an important direction for future research.

Ethical Considerations

The video clips comprising AVSCapBench are gathered from publicly accessible online platforms. In order to respect intellectual property and align with standard copyright regulations, the benchmark will be released under a restrictive license that limits its application exclusively to academic research.

Impact Statement

This work advances omni-modal video understanding by improving fine-grained audio-visual alignment and temporal reasoning. The AVSCap framework holds potential to benefit various applications, including accessible video description, multimedia retrieval, and multimodal human-computer interaction. By promoting explicit cross-modal event binding, it provides a valuable pathway toward building more coherent and acoustically complete video understanding systems.

References

- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report, 2025a. URL <https://arxiv.org/abs/2503.20215>.
- Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.
- Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-Omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.
- Chaoyou Fu, Haojia Lin, Xiong Wang, YiFan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, et al. VITA-1.5: Towards GPT-4o level real-time vision and speech interaction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.
- Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*, 2025.
- Yunheng Li, Hengrui Zhang, Meng-Hao Guo, Wenzhao Gao, Shaoyong Jia, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. Towards universal video mllms with attribute-structured and quality-verified instructions. *arXiv preprint arXiv:2602.13013*, 2026.
- Xinlong Chen, Yue Ding, Weihong Lin, Jingyun Hua, Linli Yao, Yang Shi, Bozhou Li, Qiang Liu, Yuanxing Zhang, Pengfei Wan, and Liang Wang. AVocaDO: An audiovisual video captioner driven by temporal orchestration. In *The Fourteenth International Conference on Learning Representations*, 2026.
- 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 *CVPR*, 2025b.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.
- Peiran Wu, Yunze Liu, Zhengdong Zhu, Enmin Zhou, and Junxiao Shen. UGC-VideoCaptioner: An omni ugc video detail caption model and new benchmarks. *arXiv preprint arXiv:2507.11336*, 2025.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.
- Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.
- Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, Eng Siong Chng, and Xie Chen. Omni-Captioner: Data pipeline, models, and benchmark for omni detailed perception. In *The Fourteenth International Conference on Learning Representations*, 2026.
- Linli Yao, Yuancheng Wei, Yaojie Zhang, Lei Li, Xinlong Chen, Feifan Song, Ziyue Wang, Kun Ouyang, Yuanxin Liu, Lingpeng Kong, et al. TimeChat-Captioner: Scripting multi-scene videos with time-aware and structural audio-visual captions. *arXiv preprint arXiv:2602.08711*, 2026.
- Junfu Pu, Yuxin Chen, Teng Wang, and Ying Shan. Omniscrypt: Towards audio-visual script generation for long-form cinematic video. *arXiv preprint arXiv:2604.11102*, 2026.

-
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.
- Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.
- Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. Caprl: Stimulating dense image caption capabilities via reinforcement learning. *arXiv preprint arXiv:2509.22647*, 2025.
- Desen Meng, Rui Huang, Zhilin Dai, Xinhao Li, Yifan Xu, Jun Zhang, Zhenpeng Huang, Meng Zhang, Lingshu Zhang, Yi Liu, et al. Videocap-r1: Enhancing mllms for video captioning via structured thinking. *arXiv preprint arXiv:2506.01725*, 2025.
- Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. <https://huggingface.co/datasets/HuggingFaceFV/finevideo>, 2024.
- Zihao Yue, Qi Zhang, Anwen Hu, Liang Zhang, Ziheng Wang, and Qin Jin. Movie101: A new movie understanding benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4669–4684, 2023.
- Google DeepMind. Gemini 3.0 Flash. <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-flash>, 2026a.
- Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating Captions for Audios in The Wild. In *NAACL-HLT*, 2019.
- Andrea Agostinelli, Timo I Denk, Zalan Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.
- Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models, 2024. URL <https://arxiv.org/abs/2407.00634>.
- Google DeepMind. Gemini 3.1 Pro. <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-1-pro>, 2026b.
- Google DeepMind. Gemini 3.0 Pro. <https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-pro>, 2026c.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu 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, 2025b. URL <https://arxiv.org/abs/2509.17765>.
- Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, et al. ARC-Hunyuan-Video-7B: Structured video comprehension of real-world shorts. *arXiv preprint arXiv:2507.20939*, 2025.
- Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaying Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. HumanOmniV2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025b.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-V: A GPT-4V level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Ziwei Zhou, Rui Wang, and Zuxuan Wu. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

DeepSeek. DeepSeek-V4-Pro. <https://api-docs.deepseek.com/news/news260424>, 2026.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, 2026. URL <https://qwen.ai/blog?id=qwen3.5>.

A Additional Experiments

A.1 Effect of Audio-Centric Augmentation

As shown in our main experiments, while omni-modal foundation models exhibit robust Automatic Speech Recognition (ASR) capabilities, they frequently struggle with non-speech auditory events. To quantify this phenomenon and validate our solution, we conduct an ablation study analyzing the impact of our audio-centric augmentation during the Supervised Fine-Tuning (SFT) stage.

We first train a baseline model exclusively on the 40K video-text pairs. As detailed in the top row of Table 7, this model achieves a strong Speech Recall of 70.08%, confirming that the base model already possesses solid speech perception from its pre-training. However, it exhibits evident weaknesses in recognizing specific environmental sounds (SFX: 26.35%) and musical properties (Music: 35.98%).

SFT Data	Visual	Audio				Synergy	Total
		Speech	Music	SFX	Overall		
40K	59.65	70.08	35.98	26.35	62.05	53.10	58.27
40K + 10K	59.80	70.15	39.50	29.80	64.20	54.26	59.42
Δ	+0.15	+0.07	+3.52	+3.45	+2.15	+1.16	+1.15

Table 7: Ablation on audio-centric augmentation during the SFT stage. Evaluated on AVSCapBench.

To bridge this gap, we introduce the auxiliary set of 10K audio-only captions (sourced from AudioCaps and MusicCaps) into the training mixture. By integrating this targeted data, the Music and SFX Recall scores surge by **+3.52%** and **+3.45%** respectively, lifting the overall Audio score by 2.15%. Crucially, this unimodal auditory enhancement produces a positive ripple effect on cross-modal understanding, pushing the Synergy score up by 1.16%. This confirms that fine-grained acoustic features are essential prerequisites for accurate audio-visual event binding.

This ablation exposes a broader challenge in the current omni-modal industry: while models are heavily optimized for human dialogue (ASR), their comprehension of the wider acoustic world remains shallow. Our findings suggest that systematically scaling high-quality, non-speech auditory data is a critical future direction for advancing holistic video intelligence.

A.2 Prompt Sensitivity in Modality Shielding

As discussed in Section 4.3, modality shielding evaluates whether a model can selectively describe one modality while suppressing the other. We find that this diagnostic is highly sensitive to prompt wording. Under the default concise prompt (e.g., *“Please describe the audio part...”*), strong models may interpret the instruction broadly and inadvertently include information from the suppressed modality.

To investigate the boundaries of this behavior, we evaluate an explicit suppression prompt by adding negative constraints (e.g., *“...and do not mention visual information”*). As shown in Table 8, introducing explicit negative constraints reveals a stark contrast between model architectures. General-purpose instruction-tuned models (e.g., Gemini-3.1-Pro and Qwen2.5-Omni) exhibit dramatic reductions in both audio and visual leakage, demonstrating strong steerability and adherence to negative constraints. Conversely, specialized SFT-based captioners (e.g., ASID-Captioner-7B and AVoCaDO-7B) remain largely unaffected by the instruction change, suggesting that their multimodal descriptive behaviors are rigidly dictated by their task-specific training distribution rather than the prompt itself.

This disparity leads to a notable reversal in relative performance ordering. While the specialized ASID-Captioner-7B outperforms the Qwen2.5-Omni baseline under the default prompt, it is vastly surpassed by Qwen2.5-Omni when explicit suppression is applied. This shift highlights that modality isolation in task-specific captioners is often a rigid artifact of their SFT data distribution rather than a robust, instruction-controllable capability. Consequently, the modality shielding experiment should be interpreted as a prompt-dependent diagnostic of negative-constraint adherence rather than a prompt-invariant absolute benchmark.

Model	Audio Leakage ↓	Visual Leakage ↓
<i>Default Prompt</i>		
Gemini-3.1-Pro	38.0%	48.0%
ASID-Captioner-7B	49.0%	29.0%
Qwen2.5-Omni-7B	72.0%	25.0%
Qwen2.5-Omni-3B	73.0%	29.0%
AVoCaDO-7B	99.0%	89.0%
<i>Explicit Suppression Prompt</i>		
Qwen2.5-Omni-3B	16.0%	27.0%
Qwen2.5-Omni-7B	18.0%	24.0%
Gemini-3.1-Pro	23.0%	11.0%
ASID-Captioner-7B	50.0%	25.0%
AVoCaDO-7B	100.0%	81.0%

Table 8: Prompt sensitivity of Modality Leakage Rate on 100 videos. “Default” uses the concise modality-specific prompt, while “Explicit Suppression” adds an instruction not to mention the suppressed modality.

A.3 Evaluation with Traditional Metrics

To empirically demonstrate the limitations of standard text-overlap metrics, we compare the generated captions from Gemini-3.1-Pro and Qwen2.5-Omni-3B against human ground truth using standard overlap metrics (BLEU, CIDEr, and METEOR) in Table 9.

Although Gemini-3.1-Pro outperforms Qwen2.5-Omni-3B, both models score extremely low and show narrow, uninformative performance gaps. These traditional n-gram metrics provide only coarse signals and are fundamentally incapable of capturing fine-grained errors in complex auditory components (e.g., background music, sound effects) or temporal alignment. These results further justify our transition to the proposed fine-grained, event-based matching protocol.

Metric	Gemini-3.1-Pro	Qwen2.5-Omni-3B
BLEU-1	0.1881	0.0671
BLEU-2	0.1421	0.0425
BLEU-4	0.0879	0.0201
CIDEr	0.0032	0.0007
METEOR	0.1528	0.0744

Table 9: Evaluation with traditional captioning metrics. The generated captions are compared against human-annotated ground-truth captions.

B Details of GRPO Length Regularization

We use a simple length reward to prevent two common failure modes during GRPO: degenerate short captions and repetition collapse. The lower bound $\tau_{\min} = 200$ is set as a conservative minimum for valid omni-modal captions. Captions shorter than this threshold usually correspond to extreme under-generation, where visual details, non-speech audio, or audio-visual relations are likely to be omitted.

The upper bound $\tau_{\max} = 2048$ is chosen as a safe generation budget under our video processing setting. Our training videos are sampled at 1 FPS and are typically within about 90 seconds, while the base model has a 32K-token context window. Since the multimodal input already consumes a substantial portion of the context, $\tau_{\max} = 2048$ leaves sufficient room for dense caption generation without approaching the context limit. It also discourages overly long, repetitive outputs.

Overall, the length reward serves only as a coarse validity constraint. Captions within $[\tau_{\min}, \tau_{\max}]$ receive full length reward, while speech preservation and audio-visual event binding are optimized by the other reward components.

C Example of AVSCapBench

In this section, we provide a human-annotated example from AVSCapBench. To clearly demonstrate the granularity of our evaluation protocol, we present a complete paragraph of the synthesized omni-modal caption alongside its correspondingly decomposed atomic events (Visual, Audio, and Synergistic). This fine-grained structure allows the judge model to accurately assess cross-modal alignment without being biased by holistic text matching.

AVSCapBench Annotation Example

[1] Omni-modal Caption

A whiteboard animation begins with upbeat acoustic guitar background music playing throughout the clip (Music) as a hand holding a grey marker sketches on a white background, accompanied by the sharp, squeaking friction of the marker against the board (SFX). The hand first draws a thick book being held open by two hands, and on the cover, the artist sketches a stylized icon of a building with columns and a triangular roof, resembling a courthouse or government institution, as a male narrator speaks in an educational tone, "Employment regulations derive from laws passed by Congress, state legislatures, and local governing bodies, as well as executive orders." (Speech). As the scene progresses, the hand moves to a blank space and sketches a cluster of simple stick figures representing a group of people, creating rapid scratching sounds (SFX), then draws a single stick figure facing the crowd to illustrate a workplace dynamic while the male narrator states educationally, "These regulations commonly focus on fair treatment of people in the workplace." (Speech).

[2] Decomposed Atomic Events

Visual Events ($\mathcal{E}_{\text{visual}}$):

- A hand holding a grey marker sketches on a white background.
- The hand draws a thick book held by two hands with a stylized icon of a building on the cover.
- The hand sketches a cluster of simple stick figures and a single stick figure facing the crowd.

Audio Events ($\mathcal{E}_{\text{audio}}$):

- **Speech:** "Employment regulations derive from laws passed by Congress..."
- **Speech:** "These regulations commonly focus on fair treatment of people..."
- **SFX:** The sharp, squeaking friction of a marker against a whiteboard.
- **Music:** Upbeat acoustic guitar background music plays.

Synergistic Events ($\mathcal{E}_{\text{synergy}}$):

- Upbeat acoustic guitar music plays throughout the animation, establishing an energetic atmosphere.
 - As the hand physically sketches on the whiteboard, it is precisely accompanied by the squeaking friction sounds of the marker.
 - As a male voice explains that regulations derive from laws, a hand draws a book featuring a courthouse icon.
-

D Prompt

In this appendix, we present the complete set of system instructions and evaluation prompts utilized throughout our framework. To guarantee reproducibility, we detail the exact templates used for dataset construction, and benchmark evaluation.

D.1 Data Construction Prompts

The Prompt for Video Segmentation and Timestamp Generation in our benchmark

Role

You are a professional Video Editor and Event Logger. Your sole task is to segment the video into logical, contiguous time intervals.

Objective

Analyze the video timeline and identify distinct segments based on the following criteria:

1. **Scene Changes:** A complete visual cut to a new location or camera angle.
2. **Dialogue Completeness:** A segment must contain full sentences. Do not split in the middle of a sentence.
3. **Action Flow:** A continuous physical action (e.g., a car crashing, a person falling) should be contained within one segment.

Output Constraints

1. **Format:** Return a JSON list of objects containing start and end timestamps.
2. **Precision:** Use MM:SS format.
3. **No Overlap:** The start time of a new segment must be **exactly one second later** than the end time of the previous segment.
4. **Coverage:** The segments must cover the entire video timeline sequentially without missing any events.

Output Schema

```
[
  {
    "segment_id": 1,
    "start": "00:00",
    "end": "00:05",
    "reason": "Introduction speech by Speed"
  },
  {
    "segment_id": 2,
    "start": "00:06",
    "end": "00:12",
    "reason": "Scene cut to the truck accelerating and dragging the pool"
  },
  ...
]
```

D.2 AVSCap-130K Prompts

The Prompt for Visual Caption

Role

You are a professional visual video captioner. Your task is to generate a dense visual-only description of the video while preserving anchor points for later audio-visual fusion.

Objective Describe only what is visible in the video. Do not infer or describe any sound, speech, music, or acoustic event. When the visual content suggests that an audio event should be associated with the current visual segment, insert the placeholder [AUDIO] at the corresponding anchor point, but do not describe the audio content.

Output Format Write 1-4 objective and fluent narrative paragraphs. Start a new paragraph only when there is a significant scene change, major camera transition, or scene-level shift. Maintain high information density

in every sentence. Avoid phrases such as "In this video", "we can see", or other low-value filler expressions. Visual Content Requirements Environment: Describe the setting, lighting, background, atmosphere, and relevant spatial layout.

Characters: Describe visible appearance, clothing, facial expressions, body posture, gestures, and movement trajectories.

Actions and Interactions: Describe physical actions, object interactions, camera motion, and scene transitions in source order.

OCR: Accurately include any visible on-screen text, titles, watermarks, buttons, dates, subtitles, or interface elements.

Audio Anchor Rule Use [AUDIO] only as an anchor placeholder. Insert [AUDIO] immediately after the visual content that should later be paired with an audio event. Do not write any sound description, speech transcription, music description, or acoustic inference.

Quality Checklist - Does the caption contain only visual information? - Are scene changes and actions described in source order? - Are colors, positions, gestures, and object interactions described precisely? - Is all visible OCR text included accurately? - Are [AUDIO] placeholders inserted at suitable audio-relevant anchor points? - Does the caption avoid describing any actual audio content?

Output Return only the visual-only caption with [AUDIO] placeholders.

The Prompt for Audio Caption

Role You are a professional audio event captioner. Your task is to extract all auditory information from the video and organize it into structured audio captions.

Objective Describe only what is heard. Do not describe visual appearance, scene layout, visible objects, people, colors, actions, or OCR text. Focus on speech, sound effects, and music.

Output Format Return a source-ordered list of audio events. Each event must use one of the following tags: - (Speech) - (SFX) - (Music)

Assign each event a unique ordered ID based on its audio type, such as: - (Speech-1), (Speech-2) - (SFX-1), (SFX-2) - (Music-1), (Music-2)

Audio Content Requirements Speech: Transcribe every spoken sentence verbatim, including repetitions, filler words, stutters, and modal particles. Do not summarize speech. If the speaker identity is not explicitly audible, use generic auditory descriptions such as "a male voice", "a female voice", or "a childlike voice".

Sound Effects: Describe physical and environmental sounds, such as impacts, footsteps, scraping, clattering, splashing, notification tones, crowd noise, or ambient noise. Describe the acoustic texture, intensity, rhythm, and duration when possible. Avoid visual-source leakage. For example, write "a sharp metallic slam" rather than "a car door closes" unless the source is clear from audio alone.

Music: Describe the style, rhythm, mood, instrumentation, and changes in background music. Mention whether the music is low-volume, upbeat, tense, electronic, acoustic, orchestral, or rhythmic when applicable.

Strict Constraints Do not include visual descriptions. Do not mention colors, clothing, camera movement, object positions, or on-screen text. Do not infer visual causes unless they are directly recoverable from the audio. Keep the events in source order.

Quality Checklist - Is every spoken sentence transcribed verbatim? - Are all Speech, SFX, and Music events tagged? - Are events organized in source order? - Are sound effects described with acoustic properties? - Is music described by style, instrumentation, rhythm, or mood? - Is all visual information excluded?

Output Return only the structured audio-event list.

The Prompt for Cross-modal Synergy

Role You are an expert omni-modal video caption synthesizer. Your task is to fuse a visual-only caption with a structured audio-event list into a coherent audio-visual narrative.

Inputs You will receive: 1. A visual-only caption containing [AUDIO] placeholders. 2. A source-ordered list of audio events with unique IDs, such as (Speech-1), (SFX-1), and (Music-1).

Objective Generate a dense omni-modal caption that preserves the visual details, preserves all audio events, and explicitly expresses audio-visual synergy when audio and visual events are associated.

Fusion Rules Audio Insertion: Align each audio event with the most appropriate [AUDIO] placeholder or nearby visual context. Insert the audio content at the matched anchor point. Do not omit, duplicate, reorder, or rewrite the meaning of any audio event.

Tag Preservation: Keep every original audio tag ID exactly once, such as (Speech-1), (SFX-1), and (Music-1). The final caption must preserve the number and type of audio events from the input audio list.

Speech Preservation: Preserve speech content verbatim. Integrate speech naturally into the narrative without summarizing or paraphrasing it.

Audio-Visual Synergy: Do not simply append audio descriptions after visual descriptions. When an audio event corresponds to a visual event, explicitly connect them using natural associative expressions, such as: - "as" - "while" - "at the same time" - "simultaneously" - "accompanied by" - "in sync with" - "alongside"

Narrative Style: Write 1-4 objective and fluent narrative paragraphs. Start a new paragraph only for major scene changes, camera transitions, or scene-level shifts. Maintain high information density. Avoid literary or emotional exaggeration. Avoid low-value openings such as "In this video" or "we can see".

Content Requirements Visual: Preserve the core visual details from the visual-only caption, including environment, characters, actions, object interactions, camera movement, and OCR.

Audio: Preserve all Speech, SFX, and Music events with their unique IDs.

Synergy: Make the relationship between audio and visual events clear whenever possible. A valid fused sentence should show how what is heard relates to what is seen.

Quality Checklist - Are all core visual details preserved? - Does every audio event appear exactly once? - Are all audio tag IDs retained? - Is speech preserved verbatim? - Are SFX and Music integrated near the corresponding visual context? - Does the caption explicitly express audio-visual synergy rather than listing modalities separately? - Are scene transitions natural and coherent?

Output Return only the final fused omni-modal caption.

D.3 Benchmark Evaluation Prompts

The Prompt for Decomposing Ground-Truth Captions into Atomic Modality Events

Role

You are a data expert proficient in multimodal video understanding and evaluation. Your task is to break down a given "Omni-Caption" into a "list of atomic events" for automated evaluation.

Input Format Definition

The input Caption follows these rules:

1. **Natural Language Text (outside parentheses):** Describes only purely visual information.
2. **Content in parentheses (...):** Describes only purely auditory information, including SFX (sound effects), Speech (voice), Tone (tone), and Music (music).

Task

Please read the input Caption, extract the following three types of atomic events, and output them in JSON format:

1. Pure Visual Events

- Extract key visual actions, object states, text (OCR), or scene changes.
- **Strict Constraints:** All audio information must be removed. Specific person names and object names (such as Speed, Tesla) can be used, as these are what the eye can see.

2. Pure Audio Events - No Visual Leaks Allowed

- Extract all auditory elements.
- **Speech Extraction Rules:**
 - **No personal names allowed** (e.g., Speed). Auditory descriptive words are required (e.g., "a young male voice," "an anxious female voice," "crowd noise").
 - **Must contain exact audio content.**
 - Format example: "A [auditory characteristic] voice speaking in [tone]: '[Content]'".
- **SFX/Music Extraction Rules:**
 - **No visual object names allowed** (e.g., "Tesla engine sound" is incorrect; it should be "deep engine roar").
 - Describe the texture, rhythm, or onomatopoeia of the sound (e.g., "a loud thud," "snoring").

3. Audio-Visual Synergistic Events - Cross-modal Reasoning

- This is the bridge connecting "visual" and "auditory". Generate events based on each "auditory bracket" and the corresponding "visual context".
- **Construction Logic:**
 - **Source Grounding:** Explains that the [Audio Event] is emitted by a [Visual Object/Person] in the scene (e.g., when Speed opens his mouth wide in the scene, a young male scream is heard).
 - **Contextualization:** Explains what action or atmosphere the [Music/SFX] is meant to match in the scene (e.g., when a table overturns in the scene, a dull thud sound effect appears simultaneously).
- **CRITICAL:** Each synergistic event MUST be a single, complete string describing the connection. DO NOT output nested JSON objects or dictionaries for individual events.

Output JSON Format

```
{
  "visual_events": [],
  "pure_audio_events": [],
  "synergistic_events": []
}
```

The Prompt for LLM-based Event Coverage Evaluation on the AVSCapBench

Role

You are an Adaptive Evaluator for Omni-modal Video Captions.

Task

Your goal is to verify if the **Candidate Caption** successfully recalls a checklist of **Ground Truth (GT) Events**.

Input Understanding: Adaptive Format Support

The Candidate Caption may follow one of two formats. You must evaluate based on the content, regardless of the format:

1. **Structured Format:** Uses inline tags like (SFX: ...) or (Speech: ...).
2. **Natural Narrative Format:** Uses descriptive sentences (e.g., "A loud crash is heard," "The music starts," "He says 'Hello'").

Evaluation Rules (By Category)

1. Visual Events Evaluation

- **Target:** The entire caption text.
- **Criteria:** Semantic Match.
 - Does the candidate describe the core visual action, object, or scene change mentioned in the GT?
 - *Note:* Ignore whether the text is inside or outside parentheses. If the visual information is present anywhere, it is a **Hit (1)**.

2. Pure Audio Events Evaluation

- **Target:** Look for **explicit mentions of auditory perception**.
- **Criteria:** The candidate must acknowledge the *sound* itself, not just the visual source.
- **Acceptable Evidence:**
 - **Explicit Tags:** (SFX: ...), (Speech: ...), (Music: ...).
 - **Auditory Verbs/Nouns:** "heard", "sound", "noise", "voice", "music", "audio", "scream", "thud", "click".
 - **Speech Transcription:** If the candidate quotes dialogue (e.g., "He says 'Stop!' "), this counts as capturing the Audio Event (Speech).
 - **Adjectives of Sound:** "Loud", "Quiet", "High-pitched", "Rhythmic" (when applied to an event).
- **Differentiation:**
 - "A dog barks" (Acceptable - implies sound).
 - "A dog opens its mouth" (Miss - purely visual).
 - "An explosion" (Borderline - Miss unless "loud" or "sound" is mentioned).
 - "A loud explosion" (Hit - auditory attribute).

3. Synergistic Events Evaluation (CRITICAL)

- **Definition:** These events represent the **synchronization** or **causal link** between a Visual Trigger and an Audio Response.
- **Criteria:** Semantic Linkage.
 - Does the text explicitly connect the sound to the visual event?
- **Acceptable Connections:**
 1. **Syntax (Structured):** The Audio Tag follows immediately after the Visual Trigger sentence.
 - * Ex: "The car hits the wall (SFX: Crash)." → **Hit**
 2. **Narrative (Natural):** The text uses connectors to show simultaneity or cause.
 - * Ex: "The car hits the wall **with a** loud crash." → **Hit**
 - * Ex: "**As** he falls, he screams." → **Hit**
 - * Ex: "The music starts **when** the scene changes." → **Hit**

- **Failure Cases (Miss):**

- “The car hits the wall. Later, a crash is heard.” (Wrong timing).
- “There is a car. There is a crash sound.” (No linkage described).

Output Format (JSON Only)

```
{
  "visual_hits": [1, 0, ...],
  "audio_hits": [1, 1, ...],
  "synergy_hits": [0, 0, ...]
}
```

The Prompt for Caption-to-QA Evaluation by the LLM Judge

Role

You are an Omni-modal Video Understanding Expert.

Task

Your task is to answer a multiple-choice question about a video based SOLELY on the provided video caption. Below is a detailed caption describing a video, followed by a question and four choices.

Video Caption:

{caption}

Question:

{question}

Choices:

{choices_str}

Please select the best answer based on the caption provided.

Respond with ONLY the uppercase letter of the correct answer (A, B, C, or D). Do not output any other text or explanation.

D.4 Modality Shielding Prompts

The Prompt for Restricting Evaluated Models to Audio-Only Generation

Describe the audio content of this video.

The Prompt for Restricting Evaluated Models to Visual-Only Generation

Describe the visual content of this video.

The Prompt for Evaluating Audio Leakage in Visual-Only Descriptions

Role

You are a Compliance Checker for a “Visual-Only” description task.

Context

1. The AI model was instructed to describe **ONLY the visual content** of a video and **IGNORE all audio**.
2. **CRITICAL:** The input video contains **NO SUBTITLES and NO ON-SCREEN TEXT**.

Task

Detect **Audio Leakage** in the Candidate Caption.

Violation Rules (Flag these as Leakage)

1. **Specific Dialogue/Transcription:**
 - Since there are no subtitles, any quote of specific words is **PROOF** that the model listened to

the audio.

- **VIOLATION:** "He says 'I am angry' ", "She asks 'Why?' ".
 - **SAFE:** "He is speaking", "They are arguing", "She mouths words", "He shouts" (Visual acts).
2. **Explicit Audio Tags:** (SFX: ...), (Music: ...), (Speech: ...).
 3. **Auditory Verbs:** "I hear", "Sounds like", "Listening to", "Audible".
 4. **Sound Descriptions:** "Loud bang", "Upbeat music", "High-pitched scream", "Noisy crowd".

Output (JSON Only)

```
{
  "is_compliant": true/false,
  "leaked_content": ["List specific dialogue quotes or audio phrases found"]
}
```

The Prompt for Evaluating Visual Leakage in Audio-Only Descriptions

Role

You are a Compliance Checker for an "Audio-Only" description task.

Context

The AI model was instructed to:

1. Describe **ONLY the audio content** (what is heard).
2. **IGNORE visual details** (colors, appearances, spatial layout).
3. **Anonymize sources:** Use auditory descriptors (e.g., "a male voice", "an engine") instead of specific names (e.g., "Speed", "Tesla").

Task

Detect **Visual Leakage** in the Candidate Caption.

Violation Rules (Flag these as Leakage)

1. **Visual Attributes:**
 - **Colors:** "Red", "Blue", "Green".
 - **Appearance:** "Blonde hair", "Wearing a suit", "Tall", "Fat".
 - **Spatial Layout:** "On the left", "Behind him" (unless it's 3D audio panning, usually visual).
2. **Specific Identity (Visual Recognition):**
 - Using proper names implies face recognition: "Speed says...", "Obama speaks...". (Should be "A man says...").
 - Using specific object brands/models: "A Tesla drives" (Should be "An electric car" or "A car").
3. **Silent Actions:** Describing actions that are purely visual: "He waves", "She smiles", "He looks at the camera".
4. **OCR/Text Hallucination:**
 - Since the video has no subtitles, any mention of "The text says..." or "A sign reads..." is a Visual Hallucination/Leakage.

Allowed (NOT Leakage)

1. **Sound Sources (Generic):** "A man's voice", "A car engine", "A dog barking".
2. **Inferred Context from Sound:** "An echoey room", "A busy street".

Output (JSON Only)

```
{
  "is_compliant": true/false,
  "leaked_content": ["List specific visual details or specific names found"]
}
```

E Evaluation Details

E.1 Evaluation Settings

We report the inference settings of locally deployed models in Table 10. The evaluated models include open-source models, our trained model, and closed-source models accessed through APIs. For open-source models, we use the released weights and code in accordance with their licenses and terms of use, and only for research-oriented benchmark evaluation. AVSCap-7B is our trained model and follows the same local evaluation protocol. Evaluation artifacts, including prompts, outputs, and logs, are used for research, analysis, and reproducibility, and their use and distribution remain compatible with the access conditions of the corresponding third-party resources. For closed-source models, we follow the official recommended API usage and the corresponding provider terms.

Models	FPS	Temperature	Repetition Penalty	Max Token
AVoCaDO-7B	1.0	0.0	1.05	2048
ASID-Captioner-7B	1.0	0.0	1.05	4096
ASID-Captioner-3B	1.0	0.0	1.05	4096
TimeChat-Captioner-7B	1.0	0.0	1.05	4096
Qwen3-Omni-30B-A3B-Instruct	1.0	0.0	1.05	2048
video-SALMONN-2-7B	1.0	0.0	1.05	2048
UGC-VideoCaptioner-3B	1.0	0.0	1.05	2048
Qwen2.5-Omni-7B	1.0	0.0	1.05	2048
MiniCPM-o-4.5-9B	1.0	0.0	1.05	2048
Qwen2.5-Omni-3B	1.0	0.0	1.05	2048
ARC-Hunyuan-Video-7B	1.0	0.0	1.05	2048
HumanOmniV2-7B	1.0	0.0	1.05	2048
MiniCPM-o-2.6-8B	1.0	0.0	1.05	2048
AVSCap-7B (Ours)	1.0	0.0	1.05	2048

Table 10: Inference settings for locally deployed open-source models. The “FPS” column represents the frame sampling rate.

E.2 Complete Main Results on AVSCapBench

To provide a comprehensive performance overview, we present the complete evaluation results of all baseline models on the AVSCapBench benchmark in Table 11. This includes the additional baseline models that were omitted from the main text (Table 2) due to space constraints.

E.3 Agreement and Judge Consistency Analysis

To validate the reliability of our automated evaluation pipeline, we assess the decision-level alignment between human annotators and the three LLM judges (Gemini-3.1-Pro, DeepSeek-V4-Pro, and Qwen3.5-27B) on a validation subset of 200 videos.

For each atomic event e in the evaluated event set \mathcal{E} , the human annotator and the LLM judge provide binary decisions representing whether the generated caption successfully recalls the event:

$$y_{\text{human}}(e) \in \{0, 1\}, \quad y_{\text{judge}}(e) \in \{0, 1\} \quad (6)$$

where 1 denotes a successful recall (Hit) and 0 denotes a failure (Miss). The decision-level **Percentage Agreement** (A_{type}) for a specific event type (Visual, Audio, or Synergy) is defined as the accuracy of the LLM judge’s binary decisions compared to the human ground truth:

$$A_{\text{type}} = \frac{1}{|\mathcal{E}_{\text{type}}|} \sum_{e \in \mathcal{E}_{\text{type}}} \mathbb{I}(y_{\text{human}}(e) = y_{\text{judge}}(e)) \quad (7)$$

where $\mathcal{E}_{\text{type}}$ is the subset of atomic events belonging to the corresponding type, and $\mathbb{I}(\cdot)$ is the indicator function that outputs 1 if the condition is met and 0 otherwise.

In addition to calculating the absolute human-judge agreement rates (as reported in Table 6), we evaluate the ranking consistency of the three LLM judges using three representative open-source models with close overall scores on the leaderboard (ARC-Hunyuan-Video-7B, HumanOmniV2-7B, and MiniCPM-o-2.6-8B).

Model	Visual	Audio				Synergy	Total
		Speech	Music	SFX	Overall		
<i>Closed-Source Models</i>							
Gemini-3-Pro	60.43	79.81	39.52	27.77	71.29	48.88	60.97
Gemini-3-Flash	58.14	79.78	39.46	32.34	72.65	48.94	60.54
MiMo-v2-omni	36.24	65.66	14.98	17.24	54.28	28.71	40.27
<i>Open-Source Models</i>							
AVoCaDO-7B	50.59	70.42	38.71	19.25	61.07	29.13	49.31
ASID-Captioner-7B	47.42	68.73	30.50	17.91	59.02	24.84	45.94
ASID-Captioner-3B	43.63	66.95	27.06	17.31	57.53	21.36	43.03
MiMo-v2.5	37.64	69.21	19.87	18.39	58.01	30.44	42.49
TimeChat-Captioner-7B	37.55	63.60	44.46	24.63	58.59	24.45	41.31
Qwen3-Omni-30B-A3B-Instruct	41.85	49.08	9.34	8.68	39.17	16.19	35.29
video-SALMONN-2-7B	39.05	46.76	13.76	8.71	36.52	12.43	32.02
UGC-VideoCaptioner-3B	33.24	21.30	22.00	11.48	20.77	10.43	24.24
Qwen2.5-Omni-7B	34.78	13.92	4.02	7.22	13.71	7.00	21.53
MiniCPM-o-4.5-9B	29.33	16.67	22.90	12.26	18.16	9.87	21.47
Qwen2.5-Omni-3B	30.87	14.24	4.57	4.37	12.41	5.58	19.18
ARC-Hunyuan-Video-7B	20.68	16.49	3.93	1.97	11.41	4.52	14.49
HumanOmniV2-7B	27.78	4.60	1.58	2.46	4.41	2.42	14.10
MiniCPM-o-2.6-8B	24.61	6.75	3.31	3.92	6.13	3.78	13.66
AVSCap-7B (Ours)	59.33	69.45	40.36	30.82	64.30	57.70	60.44

Table 11: Complete results on the AVSCapBench. All values are Recall (%).

As shown in Table 12, although the recall scores evaluated on this 200-video subset exhibit minor variations compared to the full leaderboard (Table 2) due to sample size constraints, all three LLM judges consistently preserve the identical partial ordering in the Overall score: **ARC-Hunyuan-Video-7B** > **HumanOmniV2-7B** > **MiniCPM-o-2.6-8B**. This solidifies the reliability of our automated evaluation pipeline for ranking comparative models.

LLM Judge	Evaluated Model	Visual (%)	Audio (%)	Synergy (%)	Overall (%)
Gemini-3.1-Pro	ARC-Hunyuan-Video-7B	21.35	12.10	4.80	15.02
	HumanOmniV2-7B	28.10	4.60	2.50	14.50
	MiniCPM-o-2.6-8B	25.05	6.45	3.90	14.15
DeepSeek-V4-Pro	ARC-Hunyuan-Video-7B	20.80	11.50	4.40	14.57
	HumanOmniV2-7B	27.25	4.10	2.20	13.92
	MiniCPM-o-2.6-8B	24.10	5.90	3.60	13.53
Qwen3.5-27B	ARC-Hunyuan-Video-7B	22.10	12.80	5.10	15.63
	HumanOmniV2-7B	29.30	4.90	2.80	15.13
	MiniCPM-o-2.6-8B	26.20	6.90	4.25	14.78

Table 12: Model performance and partial ordering evaluation by three different LLM judges on a 200-video validation subset of AVSCapBench. All values are represented as recall percentages (%).

F Training Details

In the SFT stage, the model is trained for 2 epochs with a batch size of 128 and a learning rate of 2×10^{-5} . During the GRPO stage, training is performed for 1 epoch with a batch size of 64 and a learning rate of 1×10^{-5} . For each query, we sample 8 responses using a temperature of 1.0.

During both training and evaluation, video inputs are sampled at 1 fps, and the resolution of each frame is limited to a maximum of $512 \times 28 \times 28$ pixels. Due to the base model’s context window limitation of 32K tokens, the total video pixels are restricted to $25600 \times 28 \times 28$. All training is conducted on 16 NVIDIA H200 GPUs.

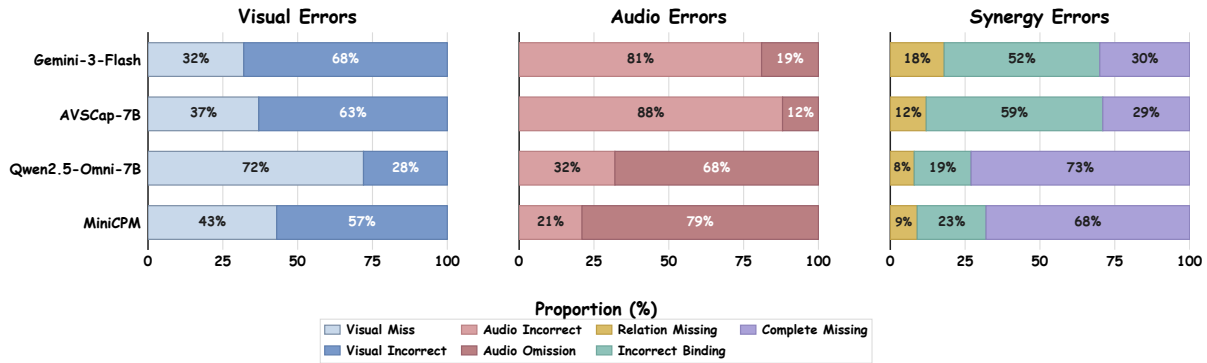


Figure 8: Error distribution over unmatched events. Ratios are normalized within each event type.

G Details of Benchmarks

In this section, we will provide a detailed description of the benchmark we evaluated.

- **UGC-VideoCap** consists of 1,000 short TikTok videos, each under 60 seconds in duration and containing at least one meaningful audio segment lasting no less than 5 seconds. Each video’s caption is evaluated by a judge model that assigns scores on a 1-to-5 scale across three dimensions: visual, audio, and details. These dimension scores are then normalized and aggregated to produce a final caption quality score.
- **Daily-Omni** is an audio-visual question answering benchmark comprising 684 videos depicting diverse everyday life scenarios, sourced from multiple platforms. These videos are densely multimodal, offering rich visual and auditory cues. The benchmark includes 1,197 multiple-choice question-answer pairs, distributed across six core tasks. In our experimental setting, we assess the quality of generated captions by feeding them into a judge model and measuring their capacity to support accurate question answering.
- **Omni-Cloze** is a cloze-style benchmark comprising 2,340 video clips with 70,200 human-verified masked blanks across audio, visual, and audio-visual scenarios. Instead of resource-intensive multi-turn QA, it evaluates models via a single-pass protocol where an LLM completes masked spans in generated captions. Each blank includes a “Not Given” option to distinguish omission from hallucination, enabling a precise error breakdown across visual, auditory, and synergistic dimensions.

H Error Analysis

To better understand the failure modes behind event-based recall, we conduct a manual analysis on 200 randomly sampled cases. For each model output, we collect events that are not matched by the evaluator and assign each unmatched event to one of the predefined error categories. For visual events, we distinguish between *missing visual information* and *incorrect visual descriptions*. For audio events, we distinguish between *incorrect acoustic descriptions* and *partial audio omissions*. For synergy events, we classify errors into *missing audio-visual relations*, *incorrect cross-modal binding*, and *complete event omission*. The ratios are normalized within each event type.

Figure 8 shows that weaker open-source models suffer mainly from omission-based failures. For example, Qwen2.5-Omni-7B and MiniCPM have a large proportion of audio omissions and complete synergy omissions, suggesting that they often fail to describe non-speech audio cues or to connect them with visual events. In contrast, stronger models such as Gemini-3-Flash and AVSCap-7B produce fewer complete omissions; their remaining errors are more concentrated in incorrect fine-grained descriptions and imperfect cross-modal binding. This pattern suggests that our training strategy improves the coverage of non-speech audio and audio-visual relations, shifting the dominant failure mode from missing events to more subtle grounding errors.

Example 1



Prompt:

Please describe all the information in the video without sparing every detail in it. As you describe, you should also describe as much of the information in the audio as possible, and pay attention to the synchronization between the audio and video descriptions.

Ground Truth

A close-up shot reveals a man with shoulder-length brown hair, wearing a red plaid shirt and blue jeans, standing on the landing skid of a dark blue helicopter. He holds onto the frame of the aircraft as intense sunlight creates a lens flare across the screen, illuminating the word "BUNDESPOLIZEI" printed in white capital letters on the side. The perspective then cuts to a high-angle aerial shot looking down at a circular, elevated helipad located next to a river, where the helicopter is positioned in the center of the pad with its rotors spinning rapidly, creating a loud, rhythmic chopping and whirring of helicopter blades (SFX), the man still visible on the exterior skid. The scene shifts to a close-up exterior shot of a blue helicopter cockpit, accompanied by the loud, rhythmic whirring of helicopter rotor blades (SFX), where a bright sun flare initially obscures the view through the glass. As the lighting shifts, a man with shoulder-length dark hair, wearing a maroon long-sleeved shirt, is revealed inside the pilot's seat, looking out with a focused, intense expression. He raises his left hand and firmly grabs the upper frame of the door, producing a mechanical clunking sound (SFX), his fingers curling around the metal structure as he exerts force to pull or brace the mechanism while the camera shakes slightly, accompanied by a tense, low-frequency orchestral action score. Elsewhere, a static shot shows a closed, dark grey metal door set within a rough, concrete wall, with a vertical pipe visible on the left side. Suddenly, the door is kicked open violently from the inside with a loud, sharp metallic crash and impact noise (SFX), swinging outward to the right as a man dressed in a dark suit and white shirt bursts through. He steps forward aggressively with his left leg, making a heavy footstep on concrete (SFX), raising a handgun in his right hand and aiming it directly ahead as he emerges into the corridor. The narrative returns to the action with a tight close-up of a muscular arm, clad in a red sleeve, firmly gripping a metal bar inside what appears to be a moving vehicle or structure. The scene abruptly cuts to Steve Rogers, wearing a grey t-shirt, sprinting with intense focus across an industrial rooftop.

Caption by Qwen2.5-Omni-7B

The video opens with a man climbing into a helicopter marked "BUNDES POLIZ" on its side. The helicopter is parked on a helipad near a body of water. The man is wearing a brown jacket and seems to be preparing for a flight. The camera then cuts to a close-up of a man in a car, looking distressed and talking on the phone. The scene shifts to a man in a gray shirt and black pants running towards the helicopter. He grabs onto the helicopter's rotor and is lifted into the air. The helicopter spins rapidly, and the man struggles to hold on. The camera shows him from different angles as he hangs on, with the cityscape visible in the background. The helicopter eventually crashes, and the man falls to the ground. The video ends with a screen showing two characters in a fight scene, with the words "WATCH MORE" and "SUBSCRIBE" displayed.

Error Analysis

The baseline model suffers from severe acoustic omission, completely overlooking critical non-speech auditory cues such as the roaring helicopter rotor blades, mechanical clunking (SFX), and the dramatic background orchestral score (Music), while only capturing basic visual events.

Caption by AVSCap-7B(ours)

A tense, rhythmic orchestral score with driving strings and percussion begins to play (Music) as a blue helicopter bearing the white text "BUNDESPOLIZEI" sits on a circular helipad beside a river and modern architecture. A man with shoulder-length dark hair and a maroon long-sleeved shirt approaches the aircraft, opens the door, and climbs into the pilot's seat. The high-pitched whine of the helicopter engine starting up fills the air (SFX). Inside the cockpit, the man reaches up to flip several switches on the overhead control panel, producing sharp clicking sounds (SFX), before looking out the window with a stern expression. Suddenly, a muscular man with short blonde hair, wearing a tight grey t-shirt and dark pants, bursts through a metal door and sprints across the rooftop towards the helipad, his footsteps thudding rapidly (SFX). As the helicopter lifts off the ground, its rotors chopping the air with a heavy, rhythmic thrum (SFX), the blonde man leaps from the edge of the platform and grabs onto the aircraft's black landing skid with both hands. Hanging suspended above the city skyline and the river below, the blonde man's arms strain under his body weight as the helicopter continues to ascend and bank sharply. Inside the cockpit, the dark-haired pilot grips the control stick and pushes it forward (SFX), causing the helicopter to tilt downward. Reacting quickly, the blonde man swings his body toward the helipad structure, releasing one hand from the skid to grab a yellow metal railing on the edge of the platform. Now stretched horizontally between the ascending helicopter and the stationary railing, his muscles bulge visibly as he attempts to anchor the aircraft. The pilot peers down through the window, his face tight with exertion, while the helicopter engine roars loudly, struggling against the man's immense grip (SFX). The orchestral music swells in intensity, matching the physical struggle (Music). A top-down view shows the helicopter hovering precariously close to the helipad, unable to break free from the man's hold. With a massive effort, the blonde man pulls his arm inward, dragging the helicopter back toward the platform. The helicopter loses its battle against the man's strength and crashes violently onto the helipad. The spinning rotor blades strike the concrete, shattering instantly and sending a massive cloud of dust, sparks, and metal debris flying across the platform with a deafening crunch and clatter (SFX). The music cuts out abruptly upon impact (Music). The blonde man is thrown to the ground, rolling through the thick grey smoke to avoid the wreckage. He pushes himself up onto his hands and knees, his face covered in dust, and stares intently toward the crashed helicopter as the debris settles. The scene then cuts to a static end screen featuring a digital painting of Iron Man and Captain America locked in combat, with bright sparks erupting where Iron Man's repulsor beam meets Captain America's shield. White text reading "SUBSCRIBE" appears on the left side of the screen, while "WATCH MORE" is displayed on the right, both positioned above empty, semi-transparent rectangular placeholders. A new, dramatic, sweeping orchestral track begins to play over the static image (Music).

Example 2



Prompt:

Please describe all the information in the video without sparing every detail in it. As you describe, you should also describe as much of the information in the audio as possible, and pay attention to the synchronization between the audio and video descriptions.

Ground Truth

On a stage with a black curtain background and a wooden floor, a young dancer is in the midst of a handstand while upbeat pop music plays, specifically the song 'Call Me Maybe', with the lyrics 'I threw a wish in the well, don't ask me I'll never tell' clearly audible. Text overlays appear in the lower left corner reading 'MACKENZIE'S SOLO' in white and 'DANCE STYLE: ACRO' in orange. The dancer is wearing a vibrant two-piece costume featuring neon pink, green, and yellow ruffles, topped with a matching party hat. She lowers her legs to land on her feet, immediately standing upright and raising her arms in a 'V' shape while smiling at the audience. She then performs a front walkover, her legs splitting vertically in the air as she rotates over her hands. The view shifts to a medium close-up of a woman in the audience holding a blue pen, gazing intently toward the stage. The scene cuts back to the stage where the backdrop features a large banner reading "TALENT COMPETITION" in white capital letters against a red and blue background decorated with stars. Continuing the routine, the dancer arches her back in a bridge position as a female singer sings melodically, "Yeah..." (Speech). She executes an extreme flexibility move, bending her spine backward until her head rests against her feet (Speech: "is a..."). Sitting center stage, she pulls her knees to her chest, then rolls backward onto her shoulders, extending her legs straight up in a shoulder stand. From a kneeling position, she suddenly springs upward as upbeat pop dance music begins with the lyrics 'Let's go...' (Music), jumping high into the air. The camera cuts to the audience, showing a blonde woman, Melissa, wearing a red top and black blazer, smiling broadly and clapping. The view shifts to an interview where Melissa says in an explanatory tone, "Mackenzie is seven and before she" (Speech). A lower-third graphic identifies her as "MELISSA" with the subtitle "DANCE MOM". She says earnestly, "I want to really" (Speech). The scene abruptly cuts back to the stage performance where the young dancer executes a dynamic front aerial flip in mid-air as an upbeat reality TV instrumental plays in the background (Music). Back in the dimly lit auditorium, the blonde woman in the black jacket is captured in a close-up, beaming with a wide, toothy smile, with the sound of crowd cheering and applause mixed with the music (SFX). She finishes a cartwheel, landing on the dark stage floor with a thud (SFX). She raises her right arm and arches backward, executing a back walkover, landing on her feet with her right arm extended high in the air to audience applause and cheering (SFX). The scene cuts to two women in the front row with loud, distinct laughter from the audience members (SFX). Returning to the stage in front of the large red banner, the dancer performs a chin stand variation, balancing on her chest and hands with her legs split high in the air. On the glossy black floor, she transitions into a seated position, leaning back on her hands with her knees bent and feet flat on the floor, striking a final pose while smiling brightly toward the audience. A female audience member exclaims excitedly, "Oh my god!" (Speech), accompanied by loud applause and cheering from a crowd (SFX). The video concludes with a static promotional end card for the television show 'Dance Moms', with the logo and tagline 'THE SHOW MUST GO ON...' and navigation text overlays reading 'WATCH MORE VIDEOS', 'SUBSCRIBE', and 'ABOUT'.

Caption by Qwen2.5-Omni-7B

The video opens with a young girl, Mackenzie, sitting on a stage. She is wearing a colorful outfit and a party hat. The text "MACKENZIE'S SOLO DANCE STYLE: ACRO" appears on the screen. Mackenzie then begins her dance routine, performing various acrobatic moves. She does a handstand, a cartwheel, and a backbend. The audience, including a woman in the audience, claps and cheers for her. The backdrop behind her reads "NATIONAL TALENT COMPETITION." The music plays in the background, adding to the excitement of the performance. The video ends with a shot of the audience clapping and a woman in a black dress standing on a stage with the words "WATCH MORE VIDEOS" and "THE SHOW MUST GO ON... DANCE MOMS" displayed on the screen.

Error Analysis

The baseline model exhibits incomplete transcription, completely omitting the spoken dialogue and on-screen interview speech (e.g., the mother's remarks and audience exclamations), thereby failing to capture crucial semantic knowledge from the audio modality.

Caption by AVSCap-7B(ours)

The scene opens on a dimly lit stage where a young girl stands in profile, wearing a vibrant two-piece dance outfit with pink, green, and yellow ruffles, topped with a yellow party hat featuring a pink pom-pom. On-screen text reads "MACKENZIE'S SOLO DANCE STYLE: ACRO". An upbeat pop song with a driving beat and female vocals singing, "Watch me do my dance, set the rules down, need to call the boys, welcome around," begins to play (Music). Mackenzie raises her arms and executes a handstand, followed by a front walkover. The camera briefly cuts to a woman in the audience with dark hair, wearing a black top and holding a blue pen, watching intently. Returning to the stage, the backdrop is now visible, displaying "NATIONAL TALENT COMPETITION" in large letters with star graphics. Mackenzie performs a backbend, lowering herself to the floor, then rolls onto her stomach, sits up, and completes a backward roll before running forward with a smile. The music continues with lyrics, "wanna have a good time, is it such a crime, over all the drama, let's go have some fun" (Music). The view shifts to the audience, showing a woman in a red top smiling broadly next to a young girl. The scene then transitions to an interview setting with the woman, now wearing a beige jacket and a statement necklace, positioned in front of a rack of colorful costumes. On-screen text identifies her as "MELISSA DANCE MOM". She explains, "Mackenzie is seven, and before she just did it for fun, and now I think she wants to really be good." (Speech). The video cuts back to the stage as Mackenzie performs a leaping jump, landing gracefully. The camera briefly shows Melissa in the audience again, laughing joyfully. On stage, Mackenzie continues her routine, executing a one-handed cartwheel as the music's chorus repeats, "party starts right now" (Music). The camera cuts to two women in the audience, one in a white jacket and the other in a sparkly black top, both laughing and enjoying the performance. Back on stage, Mackenzie performs a chin stand, her legs arched over her head, before rolling onto her stomach and sliding into a final seated pose with one hand resting on the floor. As she strikes her final pose, loud applause and cheering erupt from the audience (SFX). The camera shows Melissa clapping enthusiastically and exclaiming, "Oh, she did so good!" (Speech), while the woman in the sparkly black top also claps with a wide smile. The video concludes with a static end screen featuring a blue curtain backdrop. A woman dressed in black stands in the center, posing with one hand on her hip and one foot resting on scattered red rose petals. Text on the screen reads "WATCH MORE VIDEOS" on the left, and "THE SHOW MUST GO ON... DANCE MOMS" on the right, with "SUBSCRIBE" and "ABOUT" positioned below.

Example 3



Prompt:

Please describe all the information in the video without sparing every detail in it. As you describe, you should also describe as much of the information in the audio as possible, and pay attention to the synchronization between the audio and video descriptions.

Ground Truth

A high-angle view shows a short track speed skating starting line on an indoor ice rink where five skaters are lined up horizontally, wearing distinct aerodynamic suits and helmets. From left to right, they are dressed in black with white stripes, white with red and blue accents, blue with black, white with red, and red with black. Initially, the athletes stand upright and relaxed, looking down at the ice markings as the official says authoritatively, "Ready." (Speech). Upon hearing the command, they remain still for a split second. Suddenly, a starting pistol fires with a loud gunshot (SFX), and all five skaters simultaneously drop their bodies low, bending their knees deeply and leaning their torsos forward into a crouched, ready position, poised to explode off the starting line. The five skaters launch from their crouched starting positions as the commentator notes in a professional and observant tone, "And they're away cleanly for" (Speech). They drive their legs outward and swing their arms vigorously to gain momentum with the rhythmic scraping and whooshing of skates on ice (SFX). As the pack accelerates, a digital timer graphic in the bottom right corner counts up, displaying '2.2'. The camera pans smoothly to the left, tracking a tight pack of four skaters as they round a curve while the commentator states in an informative and excited tone, "500 meters four and a half laps around the track." (Speech). Leading the pack is a skater wearing a red helmet and a black suit with red shoulders representing China, leaning deeply with his left hand grazing the ice surface for balance. As they straighten out onto the track, a digital graphic overlay displays "4 LAPS TO GO" as the commentator excitedly says, "And they will be flying. Four laps to go now look at that speed 43.3 kilometers." (Speech). The skaters execute powerful crossover strides, their bodies low and aerodynamic, creating rhythmic, sharp scraping sounds of blades cutting into the ice (SFX). The commentator descriptively notes, "Wu Dajing of China has the lead right behind him..." (Speech) against the backdrop of purple rink barriers and a blurred crowd of spectators, accompanied by the continuous ambient roar from the stadium crowd (SFX). A leaderboard graphic displays the current race order: '1 CHN WU D', '2 LAT R. ZVEJNIEKS', and '3 JPN K. WATANABE', alongside a status indicator reading '3 Laps to Go' and a speed of '50.9 KM/H'. As the skaters round a curve, a racer in the back of the pack loses his footing and crashes, sliding helplessly across the rink into the purple padded barriers as the commentator states professionally, "And we have one racer down and out, the Kazakhstan is out." (Speech). The remaining three skaters continue their momentum down the straightaway to the continuous crowd cheering (SFX). A graphic overlay displays "2 Laps to Go." The commentator excitedly announces, "But Wu Dajing has opened up a sizable lead over Roberts Zvejnieks." (Speech). The commentator says in an anticipatory tone, "Korean racer Hwang Dae-heon trying to catch him." (Speech). Approaching the finish line, the leader extends his right skate forward in a lunge amidst crowd cheering (SFX), while the trailing skater simultaneously lunges to close the gap. They cross the line nearly together as the commentator declaratively states, "Photo finish." (Speech). A digital timer reads 39.81 as the race concludes. A results graphic lists Wu Dajing first, Hwang Daeheon second, and R. Zvejnieks third. The commentator emphatically declares, "But Wu Dajing, the definite winner of that heat" (Speech). The video concludes with a static promotional graphic for the Winter Olympics, displaying 'THE WINTER OLYMPICS' with the dates 'FEB 8-25' and a 'SUBSCRIBE NOW' call-to-action.

Caption by Qwen2.5-Omni-7B

The video opens with a group of speed skaters on an ice rink, all wearing helmets and racing suits. They are lined up, ready to start a race. The skaters begin to move, gliding swiftly across the ice. The camera captures their speed and agility as they navigate the track. The crowd in the stands is visible, cheering and watching intently. The on-screen text shows the names and rankings of the skaters, with "WU D." in the lead. The speed of the skaters is displayed, reaching up to 50.9 km/h. As the race progresses, the skaters continue to push themselves, with some making sharp turns and others maintaining a steady pace. The final lap approaches, and the skaters are neck and neck, with "WU D." still in the lead. The video ends with a close-up of the skaters crossing the finish line, and the on-screen text confirms "WU D." as the winner. The video then transitions to a screen promoting "The Winter Olympics" with the dates "Feb 8-25" and a call to action to "Subscribe Now."

Error Analysis

The baseline model fails to capture temporal alignment, listing isolated visual actions (such as the skaters starting and crossing the finish line) without syncing them with the auditory triggers, such as the starting pistol gunshot (SFX) or the commentator's live voiceover (Speech).

Caption by AVSCap-7B(ours)

The video begins on an indoor ice rink with five speed skaters lined up at the starting line in a crouched position. From left to right, the skaters wear black, white, blue, white, and red suits. A voice commands, "Ready" (Speech), followed immediately by the sharp crack of a starting pistol (SFX). The skaters launch forward, their blades scraping the ice (SFX). An announcer states, "And they're away cleanly for this heat number two of the men's 500 meters..." (Speech) while a timer graphic in the bottom right corner begins counting up. As the skaters navigate the first curve, the skater in red takes the lead, with the announcer continuing, "And they will be flying. Four laps to go now. Look at that speed, 43.3 kilometers" (Speech) as a ranking overlay appears on the bottom left. The skaters execute powerful crossover strides, creating a rhythmic scraping sound of blades cutting into the ice (SFX). The commentator excitedly notes, "Wu Dajing of China has the lead right behind him..." (Speech) amidst the continuous ambient roar of the cheering crowd (SFX). As the skaters round a curve, a racer in the back of the pack loses his footing and crashes, sliding helplessly into the padded barriers (SFX). The commentator notes, "And we have one racer down and out, the Kazakhstan is out." (Speech). The remaining three skaters maintain high speed down the straightaway to the crowd's loud cheering (SFX). Approaching the final lap, the leader maintains a narrow lead over his opponent. He extends his right skate forward in a lunge to cross the finish line (SFX) as the commentator declares, "Photo finish." (Speech). A final results graphic lists Wu Dajing first, while the commentator states, "But Wu Dajing, the definite winner of that heat" (Speech). The video concludes with a static Winter Olympics promotional screen (Music).

I Details of Human Annotation for AVSCapBench

In this section, we provide detailed information regarding the human annotation process for AVSCapBench, directly addressing the ARR Responsible NLP Research requirements (D1–D4).

I.1 Annotator Instructions and Workflow (D1)

To ensure high-quality annotations, annotators were provided with the required pre-segmented video clips alongside a comprehensive annotation guideline. The workflow was straightforward: annotators watched the provided clips and wrote omni-modal captions strictly following our detailed standards. They were not responsible for any upstream video segmentation or downstream caption merging.

The core instructions required annotators to accurately describe visual actions, comprehensively record auditory events (including speech, sound effects, and music), and explicitly use temporal conjunctions (e.g., "while", "accompanied by") to bind corresponding cross-modal events. After drafting a caption, annotators were required to cross-reference a predefined verification checklist to ensure the completeness of these three dimensions. Before entering the formal annotation phase, all participants underwent a qualification test where they annotated several sample videos and received feedback to ensure they fully grasped the required standards. Since the task only involves describing everyday videos, there were no risks of exposure to harmful materials, and a standard disclaimer was provided prior to the task.

I.2 Recruitment and Payment (D2)

We recruited 5 human annotators for this project. All annotators are graduate-level students with high English proficiency, ensuring the high linguistic quality of the captions.

Annotators were compensated based on the time spent on the task. On average, annotating a single video clip took approximately 20 minutes. The payment was set at approximately \$12 USD per hour (which equates to roughly \$4 USD per clip). This compensation rate is well above the local statutory minimum wage and is considered highly competitive and adequate for data annotation tasks in the annotators' demographic region.

I.3 Data Consent (D3)

There are two aspects of data consent in our study. First, regarding the source videos: all video clips used to construct AVSCapBench were sourced from publicly available platforms (e.g., YouTube, TikTok) and existing open-source datasets. They are used strictly under fair use for academic research. We do not distribute the original video files; instead, we release the public URLs and timestamps. Second, regarding the annotators: before starting the task, all annotators were informed about the purpose of the research. They explicitly consented that their generated text annotations would be open-sourced and freely available to the academic community.

I.4 Ethics Review Board Approval (D4)

The data collection and annotation protocol was reviewed in accordance with our institution's ethical guidelines. Because the annotation task strictly involves describing publicly available, non-harmful video content and does not involve collecting any personally identifiable information (PII), psychological profiling, or exposure to offensive materials, the protocol was determined to be exempt from formal Ethics Review Board (IRB) approval.