ATTRIBUTE-LEVEL DATA SELF-EVOLUTION FOR AUDIOVISUAL INSTRUCTION TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained video understanding relies on detailed and reliable audiovisual captions, yet caption enrichment often introduces factual errors without equally fine-grained verification and correction. Even under caption-level verification, which checks each detailed caption as a whole, we observe a critical granularity mismatch, i.e., coarse feedback fails to pinpoint fine-grained errors and omissions, limiting targeted refinement. To address this mismatch, we introduce ASID-Verify, an attribute-centered framework for verifying and refining audiovisual captions. A verifier audits each caption against the video and available speech transcripts, returning an attribute-organized report that identifies detected errors and omissions and explains each issue. A refiner uses this feedback for targeted correction and completion, and the revised caption is re-audited in the next round, enabling data self-evolution without updating the annotation models. A single round of attribute-level verification already reaches an error-free rate of 52.5%, matching two rounds of caption-level verification at 52.3%, and a second round raises it to 75.4%. With this framework, we construct ASID-1M, a million-scale collection of single- and all-attribute captions for 121K videos, and train ASID-Captioner on it via supervised fine-tuning. Across seven benchmarks spanning captioning, caption-based QA, and temporal grounding, ASID-Captioner improves fine-grained caption quality over open-source baselines and rivals Gemini-3-Pro on several captioning evaluations, while also following attribute-conditioned instructions more accurately. We will release ASID-1M, ASID-Verify, the evaluation suite, and model checkpoints.

1 INTRODUCTION

Multimodal large language models (MLLMs) (Zhang et al., 2023; Sun et al., 2024; Cheng et al., 2024) have advanced video understanding by integrating visual, audio, and natural language information. Nevertheless, understanding real-world videos remains challenging due to the temporal distribution and semantic diversity of fine-grained information. Recent approaches use expressive video captions for supervised fine-tuning (SFT), providing explicit audiovisual supervision that also supports downstream understanding and generation (Li et al., 2025a; Meng et al., 2025; Chen et al., 2026a; 2024a; Zhao et al., 2025). Because SFT learns directly from caption annotations, their coverage determines which audiovisual details are explicitly supervised, while annotation errors can propagate into the learned model. Therefore, improving both coverage and reliability is important for learning fine-grained audiovisual representations and generating accurate descriptions.

Existing video annotation pipelines provide structured captions and improve data quality through filtering or temporal alignment (Ju et al., 2024; Xue et al., 2025; Farré et al., 2024). However, these mechanisms do not directly provide attribute-specific feedback for correcting caption errors and omissions. Our analysis in Section 3.4 shows that richer annotations improve coverage but can also introduce incorrect or hallucinated content. When detailed captions are verified as a whole, a process we call caption-level verification, we observe a granularity mismatch. Fine-grained descriptions receive coarse feedback that often fails to localize errors and omissions, limiting targeted refinement and leaving unreliable content in the supervision. As shown in Figure 1(b), semantic attributes serve not merely to organize caption content, but to localize errors and omissions and guide targeted repair.

Building on this insight, we introduce ASID-Verify, an attribute-centered data curation framework for reliable audiovisual instruction data. The framework first integrates descriptions and aligned

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

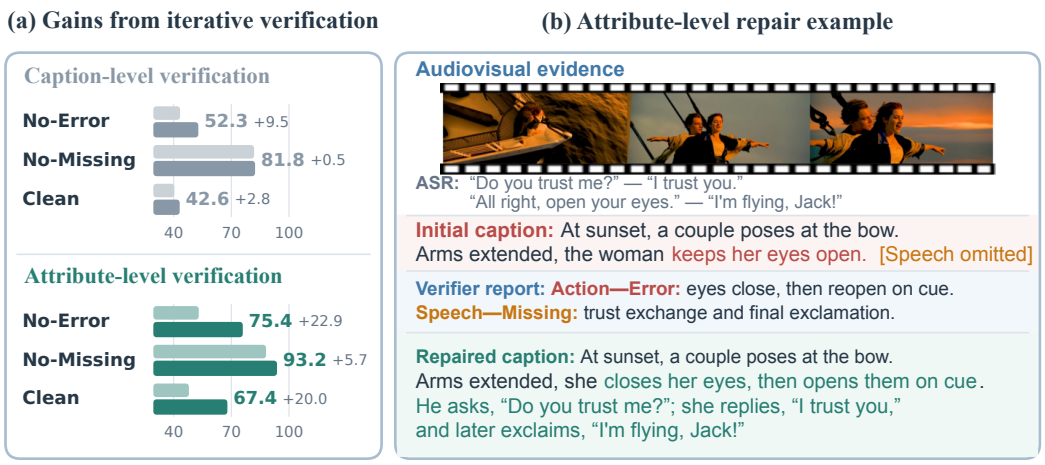


Figure 1: **Attribute-level verification improves auditing and repair.** (a) One-round (light) and two-round (dark) rates under matched budgets, with second-round gains in gray. (b) In a constructed example, the verifier's report guides the refiner's targeted edits while preserving supported content.

speech transcripts from complementary sources. A verifier then checks the resulting caption against the video and available ASR transcripts. It returns an attribute-organized report that specifies the erroneous text or missing detail and explains the corresponding evidence. The refiner uses this feedback to correct errors and add supported details while minimizing unrelated edits. The updated caption becomes the input to the next verification round. We refer to this feedback-driven annotation update as data self-evolution, with source evidence and annotation-model parameters held fixed. With this framework, we construct ASID-1M, a million-scale collection of audiovisual instructions over approximately 121K videos, each paired with eight single-attribute captions and one all-attribute caption. We then train ASID-Captioner on ASID-1M via SFT to evaluate the resulting supervision.

Figure 1(a) shows that, under matched verification budgets, attribute-level verification improves annotation accuracy and coverage over caption-level verification, with larger gains from the second round. We evaluate ASID-Captioner on seven benchmarks covering captioning, caption-based QA, and temporal grounding, together with attribute-conditioned instruction-following tests. Results show improved fine-grained caption quality and downstream performance over the evaluated open-source baselines, with the 7B model competitive with Gemini-3-Pro on several captioning evaluations. Our contributions are summarized as follows:

- ASID-Verify: An attribute-centered verification and refinement framework that addresses the granularity mismatch through localized feedback and iterative annotation updates.
- ASID-1M: A million-scale dataset of fine-grained audiovisual instructions with single- and all-attribute supervision, constructed through the proposed curation framework.
- ASID-Captioner: A video understanding model trained on ASID-1M via SFT, with improvements in fine-grained captioning and attribute-conditioned instruction following demonstrated through comprehensive evaluation.

2 Related Work

2.1 Audiovisual Multimodal Models

Video-language models support representation learning, instruction following, and conversational understanding over video content (Maaz et al., 2024; Lin et al., 2024; Wang et al., 2022; Song et al., 2024). Audiovisual approaches further incorporate speech and environmental sounds, providing complementary information for recognizing events and interpreting their context (Sun et al., 2024; Tang et al., 2025; Zhang et al., 2023; Cheng et al., 2024; Wu et al., 2025). Captioning models commonly use supervised fine-tuning to learn detailed descriptions from video-language annotations (Wang et al., 2024a; Yuan et al., 2025; Chai et al., 2025). Beyond supervised learning, several approaches explore

reinforcement learning or preference-based optimization to further refine model responses (Li et al., 2025b; Chen et al., 2026b;a; Li et al., 2026b). Our work complements these modeling approaches by improving the structure and reliability of the annotations used for instruction tuning, rather than introducing a new model architecture.

2.2 Datasets for Audiovisual Understanding

Video-text datasets provide video- and event-level supervision (Xu et al., 2016; Zhou et al., 2018; Krishna et al., 2017; Bain et al., 2021; Wang et al., 2024c; Miech et al., 2019), with subsequent efforts expanding narration and instruction annotations (Grauman et al., 2022; Maaz et al., 2024; Zhang et al., 2025; Li et al., 2026a). ShareGPT4Video combines semantic filtering with differential captioning (Chen et al., 2024a). MiraData and UltraVideo provide aspect-specific captions (Ju et al., 2024; Xue et al., 2025). UltraVideo also filters video-caption pairs by consistency, while FineVideo performs temporal alignment and anomaly filtering for audiovisual annotations (Xue et al., 2025; Farré et al., 2024). ReCorrect corrects query-interval pseudo-labels in Vid-Group (Bao et al., 2025), while Omni-Detective iteratively aggregates cross-observer evidence (Ma et al., 2026). Feedback-driven output revision has also been explored by Self-Refine (Madaan et al., 2023). Our distinction lies in using semantic attributes as explicit units for diagnosing caption errors and omissions and guiding targeted repair. Appendix D compares these pipelines in terms of filtering, temporal correction, evidence aggregation, and caption repair.

3 ASID-1M and ASID-Verify Pipeline

To construct ASID-1M, we develop ASID-Verify, a data curation pipeline that combines multi-source annotation, caption ensembling, and attribute-level verification and refinement. As shown in Figure 2, the final stage iteratively updates annotations using attribute-specific feedback.

3.1 S1: Multi-Source Annotation Generation

The pipeline begins by constructing an initial pool of approximately 135K videos from the YouTube subset of LLaVA-Video-178K (Zhang et al., 2025) and short-form videos (under 3 minutes) from FineVideo (Farré et al., 2024) to ensure broad audiovisual coverage. Initial annotations are generated using open-source audiovisual models (e.g., AVoCaDO (Chen et al., 2026a)), which jointly model visual content and audio inputs. Spoken content is transcribed and aligned via automatic speech recognition (ASR), using Whisper large-v3 (Radford et al., 2023) for transcription and WhisperX (Bain et al., 2023) for speaker-aware segmentation and utterance-level alignment. The resulting ASR subtitles, together with explicit video playback timestamps, are provided as additional inputs to closed-source multimodal models (e.g., Seed-1.6[1]) that serve as additional annotators. These sources provide complementary audiovisual descriptions and aligned speech information for constructing timestamped annotations. Thus, for each video v, we collect textual annotations $\{y_m^{(1)}(v)\}_{m=1}^{M}$ from M complementary sources, including audiovisual captions and aligned ASR transcripts.

3.2 S2: Caption Ensembling and Cross-Source Verification

Building on the textual annotations collected in Stage 1, we employ Seed-1.6 to integrate audiovisual descriptions and aligned ASR transcripts into an all-attribute draft caption, denoted as $\tilde{y}_{\mathcal{A}}^{(2)}(v)$. The integration preserves complementary details from individual sources while resolving redundancy and conflicting descriptions. However, merging multiple sources may still omit information or introduce inconsistent combinations. We therefore cross-check the draft against the source descriptions and aligned transcripts, using ASR for speech content and cross-source timestamps for temporal consistency. This stage operates on textual inputs without directly accessing the original video. Refinement then adds supported missing details and corrects conflicting statements, and the refined draft is checked again, producing $y_{\mathcal{A}}^{(2)}(v)$ for the video-grounded attribute audit in Stage 3.

[1] https://seed.bytedance.com/en/seed1_6

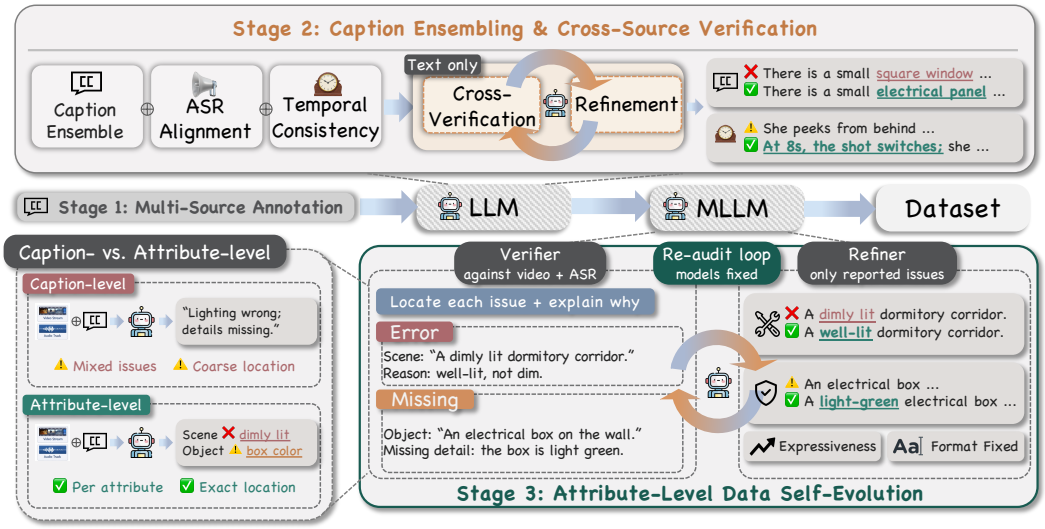
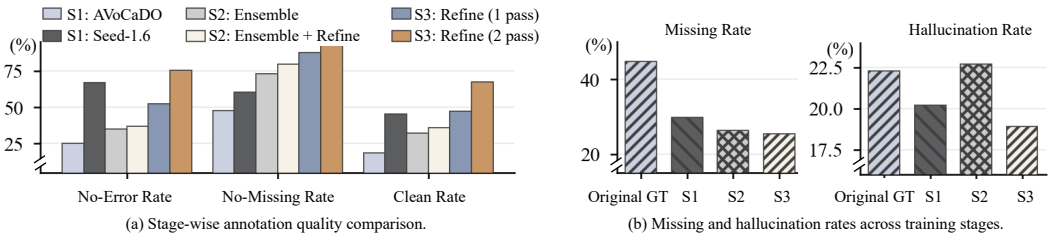


Figure 2: **Overview of ASID-Verify.** Stage 2 merges multi-source captions and verifies the result against them. Caption-level feedback mixes issues and localizes them coarsely (lower left). In Stage 3, the verifier instead audits each attribute against the video and ASR, locating and explaining each issue; the refiner fixes only the reported issues, and the result is re-audited with fixed models.



(a) Stage-wise annotation quality comparison.

(b) Missing and hallucination rates across training stages.

Figure 3: Stage-wise analysis under progressive refinement. (a) Annotation quality at each curation stage. (b) Missing and hallucination rates of models trained on captions from each stage.

3.3 S3: Attribute-Level Data Self-Evolution

Although Stage 2 integrates complementary descriptions, errors and omissions may remain in the resulting caption. We initialize the current annotation as $y_0(v) = y_{\mathcal{A}}^{(2)}(v)$ and use Seed-1.6 to audit it against the video and available ASR transcripts. A single verification call returns a structured JSON report organized by predefined semantic attributes. For each reported error or omission, the report identifies the affected text or missing detail and explains the issue. The audit also checks fluency, redundancy, and ambiguous references. At iteration t, given the audiovisual evidence $e(v)$, the verifier $\mathcal{V}$ produces feedback $f_t(v)$, and the refiner $\mathcal{R}$ updates the annotation:

$$f_t(v) = \mathcal{V}(e(v), y_t(v)), \qquad y_{t+1}(v) = \mathcal{R}(y_t(v), f_t(v)).$$

The refiner corrects errors and adds supported details while minimizing unrelated edits. The updated annotation $y_{t+1}(v)$ becomes the input to the next audit, forming the data self-evolution loop shown in Figure 2. Throughout this process, the source evidence and annotation-model parameters remain fixed. Format checks handle repetition, malformed timestamps, and invalid symbols.

Human validation. We manually inspect 1,500 sampled videos across duration ranges. Two annotators independently assess factual consistency, major-event coverage, hallucination, and temporal consistency, with disagreements resolved by a third annotator. Overall usability exceeds 98%, although minor annotation errors may remain. The detailed protocol is provided in Appendix F.1.

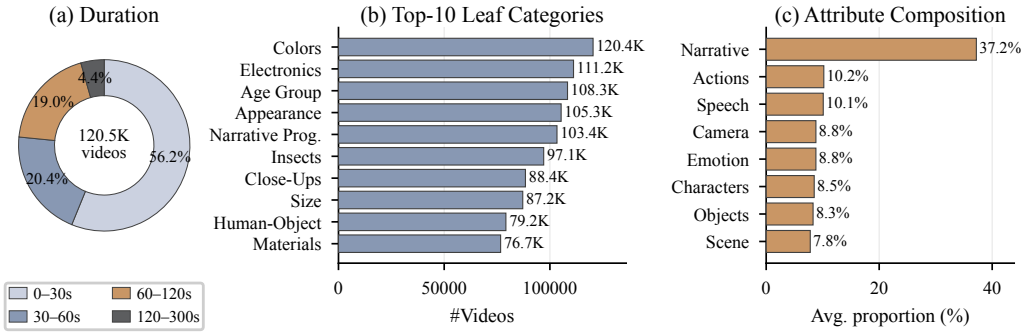


Figure 4: Dataset statistics of ASID-1M. We report the video duration distribution, the most frequent fine-grained leaf categories, and the average composition of attribute-structured captions.

3.4 Analysis of Stage-Wise Contributions

As shown in Figure 3, we assess annotation quality using No-Error, No-Missing, and Clean rates. We also evaluate downstream effects by training Qwen2.5-Omni-3B (Xu et al., 2025a) on 200K captions from different curation stages and testing on Video-SALMONN-2 (Tang et al., 2025). Multi-source ensembling improves coverage by combining complementary descriptions, but errors can remain or arise during integration. Attribute-level verification and refinement further improve annotation quality and reduce downstream missing and hallucination relative to the preceding curation stage. This motivates separating the effect of the verification unit from the benefits of multi-source ensembling.

Attribute-level versus caption-level verification. We compare the two verification units using the same pipeline and matched verification budgets at each round count. Each round includes verification followed by refinement. For No-Error, No-Missing, and Clean, attribute-level verification scores 52.5, 87.5, and 47.4 after one round and 75.4, 93.2, and 67.4 after two. Caption-level verification scores 42.8, 81.3, and 39.8 after one round and 52.3, 81.8, and 42.6 after two. Notably, attribute-level verification scores higher at both round counts, with one round achieving a No-Error score comparable to two caption-level rounds. The second round improves No-Error by 22.9 for attribute-level and 9.5 for caption-level verification. Together with the No-Missing improvements, these results support attribute-level feedback for iterative repair within the tested two-round setting.

Granularity alone is insufficient. We compare non-attribute and multi-attribute annotation before attribute-level verification. Multi-attribute annotation raises No-Missing from 42.5 to 72.9 but lowers No-Error from 59.1 to 35.2 and Clean from 40.9 to 32.3. Thus, richer descriptions improve coverage without ensuring reliability, motivating explicit verification and repair.

3.5 Dataset Statistics

ASID-1M contains approximately 121K videos, each paired with eight single-attribute captions and one all-attribute caption, yielding a million-scale instruction collection. Single-attribute captions support selective description, while all-attribute captions integrate complementary information into a coherent account. In terms of duration, Figure 4(a) shows that 56.2% of the videos are shorter than 30 seconds, while the remaining videos supply longer temporal contexts for the final training stage (Section 4). For semantic coverage, Figure 4(b) summarizes frequent fine-grained categories derived from the annotations, such as appearance, objects, interactions, and narrative content. Since a video may contain several of these categories, they describe complementary aspects rather than mutually exclusive video classes. For caption composition, Figure 4(c) shows that narrative descriptions account for the largest share of attribute text, while the other attributes provide focused descriptions of actions, speech, camera, emotion, characters, objects, and scenes.

4 Progressive Attribute Learning

For each video, ASID-1M provides supervision at two granularities, eight single-attribute captions and one all-attribute caption, both curated through ASID-Verify. We use them to train ASID-

Table 1: Model performance on the audiovisual captioning benchmarks. Evaluation is conducted using GPT-4.1 as the judge model for Video-SALMONN-2 and GPT-4o for UGC-VideoCap.

Model	Size	Modality	video-SALMONN-2 testset			UGC-VideoCap			
			Miss ↓	Hall. ↓	Total ↓	Audio ↑	Visual ↑	Detail ↑	Avg. ↑
Gemini-3-Pro	-	A + V	19.1	16.6	35.7	80.4	84.7	80.6	81.9
Gemini-2.5-Pro	-	A + V	18.1	13.3	31.3	69.5	74.7	73.7	72.6
Gemini-2.5-Flash	-	A + V	19.3	13.9	33.3	69.1	75.8	74.0	73.0
InternVL3.5 (Wang et al., 2025)	8B	V	53.8	25.5	79.4	47.9	64.8	59.5	57.4
Qwen2.5-VL (Bai et al., 2025)	7B	V	40.5	17.0	57.5	46.6	69.1	62.3	59.3
HumanOmniV2 (Yang et al., 2025a)	7B	A + V	49.2	**12.3**	61.6	45.6	66.3	59.5	57.1
ARC-Hunyuan-Video (Ge et al., 2025)	7B	A + V	45.7	12.5	58.2	52.7	56.0	55.8	54.8
Qwen2.5-Omni (Xu et al., 2025a)	7B	A + V	41.7	15.4	57.1	46.9	66.1	60.0	57.7
MiniCPM-o-2.6 (Yao et al., 2024)	8B	A + V	42.2	14.3	56.5	38.6	68.5	57.7	54.9
video-SALMONN-2 (Tang et al., 2025)	7B	A + V	21.2	17.6	38.8	61.8	71.4	68.5	67.2
UGC-VideoCaptioner (Wu et al., 2025)	3B	A + V	31.6	17.0	48.6	61.4	58.4	57.5	59.1
Qwen3-Omni-Instruct (Xu et al., 2025b)	30B-A3B	A + V	32.0	13.6	45.6	67.5	74.8	72.3	71.5
Qwen3-Omni-Captioner (Xu et al., 2025b)	30B-A3B	A + V	31.0	16.6	47.6	69.0	75.5	72.3	72.3
AVoCaDO (Chen et al., 2026a)	7B	A + V	21.1	16.2	37.3	73.0	74.6	71.8	73.2
ASID-Captioner (Ours)	3B	A + V	23.4	18.3	41.7	78.6	**84.8**	**80.2**	**81.2**
ASID-Captioner (Ours)	7B	A + V	**20.5**	15.4	**35.9**	**79.1**	84.4	**80.2**	**81.2**

Captioner from Qwen2.5-Omni (Xu et al., 2025a) for two complementary capabilities, describing a requested attribute on demand and composing all attributes into a coherent caption. The recipe first learns single attributes, then integrates them into all-attribute captions, and finally adapts to longer temporal contexts. These training stages, abbreviated as SA, AA, and LC, are distinct from the curation stages S1 to S3 in Section 3.

Training objective. Given a video v, a requested attribute set $B \subseteq \mathcal{A}$, and its target caption y_B, all stages use the same supervised objective:

$$\mathcal{L}_s = -\mathbb{E}_{(v,B,y_B)\sim\mathcal{D}_s}\left[\log p_\theta(y_B \mid v, B)\right], \qquad s \in \{\mathrm{SA}, \mathrm{AA}, \mathrm{LC}\}, \tag{1}$$

where $\mathcal{D}_s$ denotes the supervision used at stage s.

Single-Attribute Learning (SA). Each instruction requests one attribute, $B = \{a\}$, with the attribute-specific caption y_a as the target. This encourages the model to describe the requested attribute rather than always producing an unrestricted description.

All-Attribute Learning (AA). We then train on clips up to 30 seconds using the full attribute set $B = \mathcal{A}$. The target caption integrates complementary attributes into a coherent description, requiring the model to describe their relationships within the same temporal context.

Longer-Context Adaptation (LC). The final stage retains all-attribute supervision and extends training to clips of 30 seconds to three minutes. This exposes the model to descriptions spanning longer event sequences and more widely separated audiovisual information. We use LoRA for this stage, with the full training configuration provided in Table 7.

Inference. Given a video and a user-selected attribute subset, the model generates a caption conditioned on the requested content without retraining. We evaluate attribute-conditioned instruction following in the experiments. The stages above describe the recipe used for the reported checkpoints, while alternative training schedules are compared in Appendix F.4.

5 Experiments

5.1 Benchmarks

We evaluate ASID-Captioner on seven benchmarks that offer complementary perspectives on caption quality. **Audiovisual captioning:** Video-SALMONN-2 (Tang et al., 2025) measures missing and hallucinated content, while UGC-VideoCap (Wu et al., 2025) evaluates audio, visual, and descriptive detail. **Attribute-wise visual captioning:** VDC (Chai et al., 2025) reports accuracy and judge scores for camera, summary, background, main object, and detailed descriptions. **Generation-oriented caption evaluation:** VidCapBench-AE (Chen et al., 2025) assesses accuracy, precision, coverage, and conciseness across aesthetics, content, motion, and physical laws. **Caption-based QA:** Daily-Omni (Zhou et al., 2025) and WorldSense (Hong et al., 2026) assess whether captions

Table 2: Comparison of attribute-wise visual captioning performance on the VDC benchmark. Accuracy and judge scores are reported for each visual aspect, with Seed-1.6 used as the judge model.

Model	Size	Camera	Short	Background	Main Object	Detailed	Avg.
		Acc / Score ↑	Acc / Score ↑	Acc / Score ↑	Acc / Score ↑	Acc / Score ↑	Acc / Score ↑
Gemini-3-Pro	-	35.5 / 1.5	27.6 / 1.2	42.3 / 1.9	45.4 / 2.0	40.2 / 1.8	38.2 / 1.7
Gemini-2.5-Pro	-	35.0 / 1.5	26.6 / 1.1	41.5 / 1.8	42.3 / 1.8	38.5 / 1.7	36.8 / 1.6
Gemini-2.5-Flash	-	34.5 / 1.4	26.5 / 1.1	41.1 / 1.7	41.6 / 1.8	38.1 / 1.7	36.4 / 1.5
Qwen2.5-Omni (Xu et al., 2025a)	3B	28.0 / 1.3	23.2 / 1.2	35.1 / 1.6	34.6 / 1.7	32.4 / 1.4	30.7 / 1.4
Qwen2.5-Omni (Xu et al., 2025a)	7B	29.5 / 1.3	26.1 / 1.2	38.2 / 1.7	37.3 / 1.7	34.2 / 1.5	33.1 / 1.5
ARC-Hunyuan-Video (Ge et al., 2025)	7B	29.1 / 1.3	25.5 / 1.2	37.4 / 1.6	36.5 / 1.5	34.3 / 1.5	32.6 / 1.4
Qwen3-Omni-Instruct (Xu et al., 2025b)	30B-A3B	33.7 / 1.4	26.3 / 1.1	41.8 / 1.9	42.7 / 1.9	38.2 / 1.6	36.5 / 1.6
Qwen3-Omni-Captioner (Xu et al., 2025b)	30B-A3B	34.4 / 1.4	26.8 / 1.2	42.1 / 1.9	44.7 / 2.0	39.4 / 1.7	37.5 / 1.6
AVoCaDO (Chen et al., 2026a)	7B	35.3 / 1.5	27.1 / 1.2	41.0 / 1.8	40.0 / 1.8	37.9 / 1.7	36.3 / 1.6
ASID-Captioner (Ours)	3B	37.0 / 1.6	28.4 / 1.3	45.5 / 2.0	44.7 / 2.0	41.7 / 1.8	39.5 / 1.7
ASID-Captioner (Ours)	7B	**38.2 / 1.7**	**28.8 / 1.3**	**46.9 / 2.1**	**47.4 / 2.1**	**43.2 / 1.9**	**40.9 / 1.8**

Table 3: Results on the VidCapBench-AE benchmark. Acc, Pre, Cov, and Con denote accuracy, precision, coverage, and conciseness, respectively, with Con scaled by 100. Scores are computed using GPT-4o under the VidCapBench-AE evaluation protocol (Chen et al., 2025).

Model	Size	Overall	Video Aesthetics	Video Content	Video Motion	Physical Laws
		Acc↑/Pre↑/Cov↑/Con↓	Acc↑/Pre↑/Cov↑/Con↓	Acc↑/Pre↑/Cov↑/Con↓	Acc↑/Pre↑/Cov↑/Con↓	Acc↑/Pre↑/Cov↑/Con↓
Gemini-3-Pro	-	18.9 / 58.1 / 92.8 / 3.7	16.7 / 51.4 / 89.8 / 3.3	18.9 / 61.5 / 94.1 / 3.7	10.7 / 39.8 / 91.6 / 2.1	38.2 / 58.3 / 93.5 / 7.5
Gemini-1.5-Pro-002	-	17.1 / 54.8 / 87.4 / 9.2	16.4 / 47.6 / 85.4 / 8.8	16.9 / 57.8 / 88.5 / 9.1	9.8 / 45.1 / 80.9 / 5.3	28.4 / 59.3 / 88.2 / 15.3
GPT-4o-20240806	-	16.8 / 57.4 / 86.0 / 5.9	14.1 / 47.6 / 83.4 / 4.9	17.5 / 61.7 / 87.2 / 6.1	10.2 / 41.3 / 84.0 / 3.6	27.9 / 62.1 / 85.4 / 9.7
Pixtral (Agrawal et al., 2024)	124B	13.0 / 48.3 / 80.5 / 3.0	13.9 / 44.6 / 80.2 / 3.2	11.9 / 50.0 / 80.5 / 2.7	6.2 / 28.3 / 81.8 / 1.4	27.9 / 55.9 / 83.2 / 6.4
InternVL2 (Chen et al., 2024b)	76B	7.4 / 35.6 / 78.9 / **0.7**	5.8 / 27.6 / 76.2 / **0.6**	7.2 / 38.1 / 80.1 / **0.7**	4.4 / 24.4 / 80.0 / **0.4**	23.1 / 55.0 / 78.1 / **2.3**
Qwen2-VL (Wang et al., 2024b)	72B	12.2 / 46.8 / 79.0 / 7.7	12.0 / 42.5 / 79.2 / 7.6	11.5 / 48.4 / 78.8 / 7.3	5.8 / 28.6 / 77.8 / 3.7	27.1 / 59.6 / 80.9 / 17.2
Tarsier (Wang et al., 2024a)	34B	13.5 / 50.8 / 82.1 / 15.1	14.7 / 43.9 / 85.5 / 16.4	12.4 / 53.7 / 80.4 / 13.8	7.1 / 38.1 / 84.0 / 7.9	28.1 / **61.7** / 83.9 / 31.4
Aria (Li et al., 2024)	25B	14.1 / 51.5 / 84.4 / 4.5	13.0 / 44.0 / 82.7 / 4.2	13.9 / 54.9 / 85.3 / 4.4	7.1 / 34.2 / 81.8 / 2.3	27.9 / 56.8 / 83.7 / 8.9
Pixtral (Agrawal et al., 2024)	12B	11.0 / 39.5 / 79.6 / 5.2	14.5 / 42.7 / 82.8 / 6.8	8.6 / 37.9 / 78.4 / 4.0	3.6 / 18.6 / 69.3 / 1.7	28.6 / 52.4 / 82.4 / 13.5
LLaVA-NeXT-Video (Zhang et al., 2024b)	7B	10.6 / 42.3 / 79.4 / 3.9	11.3 / 39.9 / 82.2 / 4.2	9.6 / 43.2 / 78.1 / 3.5	4.4 / 23.7 / 75.1 / 1.7	24.4 / 54.5 / 82.9 / 9.0
LongVA (Zhang et al., 2024a)	7B	10.8 / 43.0 / 79.3 / 6.1	12.8 / 42.1 / 83.8 / 7.3	9.2 / 43.4 / 77.2 / 5.2	4.9 / 25.1 / 79.6 / 2.8	24.9 / 52.9 / 83.2 / 14.1
mPLUG-Owl3 (Ye et al., 2025)	7B	14.5 / 49.6 / 84.4 / 6.9	12.9 / 40.7 / 83.7 / 6.1	14.8 / 53.5 / 85.1 / 7.0	5.3 / 33.3 / 80.0 / 2.5	26.9 / 55.7 / 81.7 / 12.8
InternVL2 (Chen et al., 2024b)	8B	10.2 / 43.0 / 84.9 / 2.5	9.1 / 36.3 / 84.4 / 2.2	10.0 / 46.1 / 85.2 / 2.4	4.4 / 18.0 / 81.3 / 1.1	23.6 / 52.8 / 85.7 / 5.8
Qwen2-VL (Wang et al., 2024b)	7B	11.1 / 47.1 / 77.0 / 6.4	12.4 / 44.3 / 78.7 / 7.2	9.9 / 48.3 / 75.9 / 5.7	4.0 / 22.7 / 78.2 / 2.3	26.1 / 59.4 / 81.2 / 15.1
CogVLM2-Caption (Yang et al., 2025b)	5B	13.1 / 49.2 / 85.1 / 8.4	12.5 / 45.2 / 83.1 / 8.0	12.7 / 50.8 / 86.3 / 8.1	5.7 / 33.9 / 82.7 / 3.7	27.9 / 59.9 / 82.7 / 17.8
ASID-Captioner (Ours)	3B	**18.2** / 58.3 / 93.0 / 3.6	**15.5** / **50.7** / 90.8 / 3.1	**18.7** / 62.5 / 94.2 / 3.7	8.9 / 34.0 / **90.6** / 1.8	**33.7** / 55.6 / 89.5 / 6.7
ASID-Captioner (Ours)	7B	**18.2** / **60.0** / **93.3** / 3.7	15.4 / 49.7 / **91.2** / 3.1	**18.7** / **64.8** / **94.5** / 3.8	**12.9** / **42.7** / 88.6 / 2.6	32.5 / 60.2 / **91.7** / 6.6

preserve information needed for question answering. **Caption-based temporal grounding:** Charades-STA (Gao et al., 2017) evaluates event localization using mIoU and recall at multiple IoU thresholds. For caption-based QA and grounding, a fixed Gemini-2.5-Pro solver uses generated captions rather than videos, measuring the information preserved in the captions.

5.2 Main Results

Audiovisual captioning. Table 1 shows that ASID-Captioner improves caption coverage while maintaining competitive reliability. On Video-SALMONN-2, the 7B model reduces total error from 57.1 for Qwen2.5-Omni to 35.9, primarily by lowering missing content from 41.7 to 20.5 while maintaining the same hallucination score of 15.4. It also improves over AVoCaDO's total error of 37.3 and performs close to Gemini-3-Pro at 35.7. On UGC-VideoCap, both model sizes achieve an average score of 81.2, compared with 73.2 for AVoCaDO and 81.9 for Gemini-3-Pro. An additional evaluation with Qwen3.5-27B as judge also favors ASID-Captioner over AVoCaDO on missing and hallucination, while Gemini-3-Pro remains better on both metrics (Appendix F.3).

Attribute-wise visual captioning. Table 2 reports results across five visual aspects on VDC. The 7B model achieves an average accuracy of 40.9, compared with 33.1 for Qwen2.5-Omni, 36.3 for AVoCaDO, and 38.2 for Gemini-3-Pro. It also obtains higher accuracy than these baselines on each evaluated aspect, supporting the benefit of the supervision for fine-grained visual captioning.

Generation-oriented caption evaluation. Table 3 evaluates caption quality across video aesthetics, content, motion, and physical laws. The 7B model achieves overall accuracy, precision, and coverage of 18.2, 60.0, and 93.3, respectively. Compared with Gemini-3-Pro, it obtains higher overall precision and coverage, although its accuracy is slightly lower. For video content, the model reaches a precision of 64.8 and coverage of 94.5, indicating detailed descriptions that retain relevant semantic information. It also achieves higher accuracy on video motion than Gemini-3-Pro, while performance varies across the remaining aspects. These results support the utility of the generated captions for describing content and dynamics relevant to video generation.

Caption-grounded QA. Table 4 evaluates whether generated captions preserve sufficient information for downstream reasoning. A fixed Gemini-2.5-Pro solver answers questions using captions only. The 7B model achieves the highest accuracy among the compared models on both Daily-Omni and WorldSense, reaching 61.2 and 34.0, respectively, compared with 50.1 and 25.7 for AVoCaDO. The 3B model also surpasses AVoCaDO on both benchmarks, achieving 55.5 and 32.3. These results suggest that the generated captions retain useful audiovisual information beyond a coarse summary, supporting downstream question answering.

Table 4: Caption-based QA with Gemini-2.5-Pro.

Model	Daily-Omni↑	WorldSense↑
Gemini-2.5-Pro	60.2	33.8
Gemini-2.5-Flash	55.3	31.0
HumanOmniV2	8.2	6.6
ARC-Hunyuan-Video	8.6	8.7
MiniCPM-o-2.6	9.8	7.2
Qwen2.5-Omni	13.4	8.6
UGC-VideoCaptioner	17.0	11.2
video-SALMONN-2	29.9	18.2
Qwen3-Omni-Instruct	17.5	12.7
Qwen3-Omni-Captioner	27.2	14.1
AVoCaDO	50.1	25.7
ASID-Captioner (3B)	<u>55.5</u>	<u>32.3</u>
ASID-Captioner (7B)	**61.2**	**34.0**

Caption-based temporal grounding. Table 5 evaluates temporal grounding by asking Gemini-2.5-Pro to predict event boundaries using captions only. ASID-Captioner performs strongly across all metrics, especially at higher IoU thresholds (R1@0.5 and R1@0.7). This indicates that its captions preserve temporally precise information rather than coarse event references.

Table 5: Caption-based temporal grounding with Gemini-2.5-Pro.

Model	mIoU↑	R1@.3↑	R1@.5↑	R1@.7↑
Gemini-3-Pro	8.8	46.4	26.0	10.2
Gemini-2.5-Pro	27.6	45.7	25.6	8.7
Gemini-2.5-Flash	6.1	40.0	19.5	7.1
ASID-Captioner (3B)	24.1	38.0	20.0	8.0
ASID-Captioner (7B)	**28.5**	**44.1**	**26.3**	**11.0**

5.3 Ablation Study

The supervision-variant ablation uses the same source-video subset from LLaVA-Video-178K/FineVideo, 20K training instances, and the same optimization settings for Qwen2.5-Omni-3B, while changing the annotations. This controls source-video identity and instance count, but does not by itself equalize target-token count. The full-data training-stage ablation is reported separately.

Supervision variants for captioning. Table 6a analyzes the effect of different supervision data variants on captioning performance. Using raw annotations yields only small improvements over the baseline, indicating that unstructured ground-truth captions provide weak supervision. Non-attribute annotation improves descriptive richness, but still leaves many fine-grained details unmodeled, resulting in incomplete captions. In contrast, multi-attribute supervision achieves the best overall performance, substantially reducing missing content while improving detailed caption quality, demonstrating the benefit of structured, attribute-level supervision. It is worth noting that encouraging finer-grained descriptions may cause some previously unannotated but relevant details (e.g., on-screen text) to be penalized as hallucinations under the current evaluation protocol.

Stage-wise training ablation. Table 11 reports captioning performance after each cumulative training checkpoint. Training with SA yields limited improvement, indicating that single-attribute supervision mainly supports basic attribute grounding but is insufficient for comprehensive captioning. Adding AA substantially reduces missing content and total error while improving detailed caption accuracy, although hallucination increases from 12.8 to 19.9, exposing a coverage-reliability trade-off. LC then reduces both missing content and hallucination and achieves the strongest detailed-caption result, showing the benefit of longer-context adaptation. In the additional SA/AA schedule comparison, mixed training obtains lower missing and hallucination than SA→AA (26.4/19.3 versus 26.8/20.1). Thus, the staged recipe should not be interpreted as evidence that progressive ordering is necessary; the full comparison, including single-stage and reversed-order schedules, is provided in Appendix F.4.

5.4 Attribute-level Instruction Following

To evaluate attribute-level instruction following, we prompt each model with attribute sets containing one to four attributes, and use Gemini-2.5-Pro as an automatic judge to assess whether the generated captions explicitly satisfy the requested attributes. As shown in Table 6b, existing captioning models struggle to follow attribute-specified prompts, exhibiting low accuracy even with one or two attributes.

Table 6: Controlled analyses of supervision and attribute-level instruction following.

(a) Supervision variants.

Dataset	Video-SALMONN-2			VDC
	Miss ↓	Hall. ↓	Total ↓	Acc./Score ↑
Base model	44.8	16.4	61.2	31.9 / 1.4
Original GT	43.9	22.3	66.2	35.5 / 1.5
Non-attribute	30.6	18.5	49.1	38.5 / 1.6
Multi-attribute	**25.5**	**18.9**	**43.4**	**40.2 / 1.7**

(b) Instruction following.

Model	Size	1↑	2↑	3↑	4↑
Gemini-3-Pro	-	0	0	2.0	3.0
Gemini-2.5-Pro	-	0	0	1.5	2.5
Qwen2.5-Omni	7B	14.0	2.5	6.5	5.5
AVoCaDO	7B	0	0	0	0
ASID-Captioner (Ours)	3B	**52.3**	**63.0**	9.5	**7.5**
ASID-Captioner (Ours)	7B	47.0	58.0	**9.6**	6.5

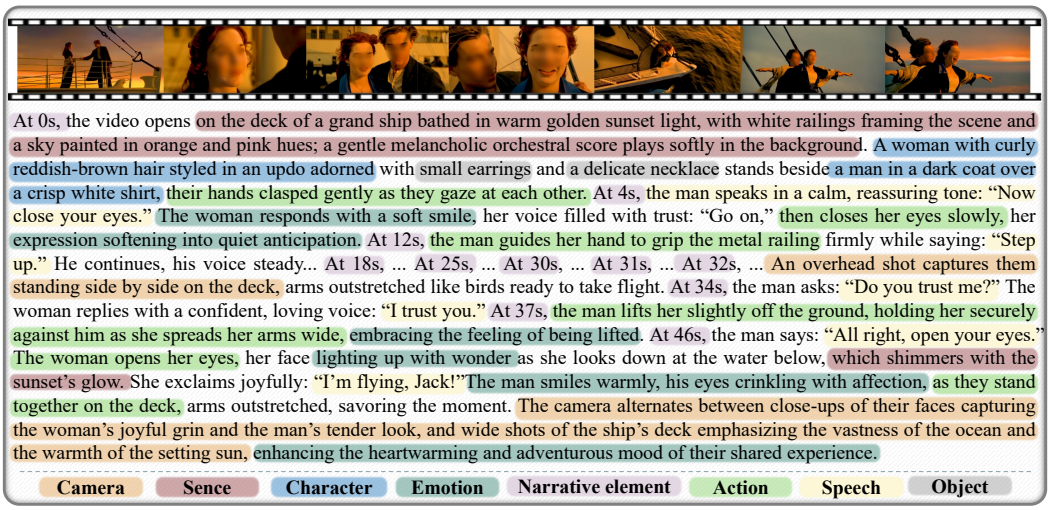


Figure 5: Example of an attribute-structured audiovisual caption generated by ASID-Captioner, with timestamps and grounded speech; color highlights indicate the corresponding attribute groups.

In contrast, ASID-Captioner consistently achieves higher instruction following accuracy. Although trained only with single-attribute and all-attribute supervision, ASID-Captioner shows improvements on intermediate attribute combinations. However, scores remain low for three and four requested attributes, indicating that reliable compositional instruction following remains challenging.

5.5 Qualitative Analysis

Figure 5 shows a representative caption generated by ASID-Captioner. The caption is explicitly timestamped and integrates multiple complementary aspects of the video, including visual content, camera cues, actions, and grounded speech, into a coherent and temporally aligned narrative. More qualitative examples are provided in Section H.

6 Conclusions

In this work, we present ASID-Verify, an attribute-centered framework for constructing reliable audiovisual instruction data. By organizing verification feedback around semantic attributes, the framework supports targeted correction and completion, with revised annotations fed into subsequent verification rounds. Our analysis shows that attribute-level verification improves annotation accuracy and coverage over caption-level verification within the evaluated one-round and two-round settings. The resulting ASID-1M provides million-scale single- and all-attribute supervision for training ASID-Captioner through SFT. Experiments demonstrate the utility of this supervision for fine-grained captioning and attribute-conditioned instruction following. These findings support attribute-level verification as a practical mechanism for audiovisual data self-evolution.

ETHICS STATEMENT

This work advances multimodal machine learning by enabling fine-grained audiovisual understanding and attribute-structured video captioning. The proposed dataset and pipeline improve the completeness and reliability of video descriptions, supporting applications such as video understanding, education, and downstream multimodal research. As with most data-driven multimodal systems, model outputs may still contain errors or be incomplete and should be interpreted with care. These risks are common to current multimodal learning approaches. This work is intended as a research contribution rather than a deployment-ready system, and we encourage responsible use in accordance with applicable ethical guidelines.

REPRODUCIBILITY STATEMENT

The dataset construction and verification procedure is described in Section 3, and the progressive training objectives are described in Section 4. The appendix reports the training configuration, annotation sources, benchmark protocols, additional results, and prompts used for automatic evaluation. Code, data, model checkpoints, and the evaluation suite will be released to support independent reproduction.

AI USE STATEMENT

Generative AI systems are integral to the research methodology in this work. As described in the main paper and appendix, we use large multimodal and language models to generate, ensemble, verify, decompose, evaluate, and refine synthetic audiovisual annotations, and as automatic judges or downstream solvers in several experiments. Generative AI also assisted with manuscript editing, figure and table preparation, and bibliography formatting, but was not used to generate or modify experimental results, and all numbers reported in figures and tables come from our experiments. The authors reviewed all AI-assisted changes, verified every reference against its original publication, and take full responsibility for the final content of this work.

REFERENCES

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Int. Conf. Comput. Vis.*, pp. 1708–1718, 2021.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*, 2023.

Peijun Bao, Chenqi Kong, Siyuan Yang, Zihao Shao, Xinghao Jiang, Boon Poh Ng, Meng Hwa Er, and Alex Kot. Vid-group: Temporal video grounding pretraining from unlabeled videos in the wild. In *Int. Conf. Comput. Vis.*, pp. 20541–20550, 2025.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. In *Int. Conf. Learn. Represent.*, pp. 16845–16889, 2025.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. In *Adv. Neural Inform. Process. Syst.*, volume 37, pp. 19472–19495, 2024a.

Xinlong Chen, Yuanxing Zhang, Chongling Rao, Yushuo Guan, Jiaheng Liu, Fuzheng Zhang, Chengru Song, Qiang Liu, Di Zhang, and Tieniu Tan. Vidcapbench: A comprehensive benchmark of video captioning for controllable text-to-video generation. In *Annu. Meet. Assoc. Comput. Linguist.*, pp. 8543–8563, 2025.

Xinlong Chen, Yue Ding, Weihong Lin, Jingyun Hua, Linli Yao, Yang Shi, Bozhou Li, Qiang Liu, Yuanxing Zhang, Pengfei Wan, et al. Avocado: An audiovisual video captioner driven by temporal orchestration. In *Int. Conf. Learn. Represent.*, pp. 24338–24362, 2026a.

Xinlong Chen, Yuanxing Zhang, Yushuo Guan, Weihong Lin, Zekun Wang, Bohan Zeng, Yang Shi, Sihan Yang, Qiang Liu, Pengfei Wan, et al. Vidbridge-r1: Bridging qa and captioning for rl-based video understanding models with intermediate proxy tasks. In *Int. Conf. Learn. Represent.*, pp. 116434–116461, 2026b.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Sci. China Inf. Sci.*, 67(12):220101, 2024b.

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.

Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Int. Conf. Comput. Vis.*, pp. 5277–5285, 2017.

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.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 18995–19012, 2022.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *Int. Conf. Learn. Represent.*, pp. 52423–52443, 2026.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. In *Adv. Neural Inform. Process. Syst.*, volume 37, pp. 48955–48970, 2024.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Int. Conf. Comput. Vis.*, pp. 706–715, 2017.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025a.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, et al. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024.

Shihao Li, Yuanxing Zhang, Jiangtao Wu, Zhide Lei, Yiwen He, Runzhe Wen, Chenxi Liao, Chengkang Jiang, An Ping, Shuo Gao, et al. If-vidcap: Can video caption models follow instructions? In *Int. Conf. Learn. Represent.*, pp. 23157–23201, 2026a.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025b.

Yunheng Li, Jing Cheng, Shaoyong Jia, Hangyi Kuang, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. Tempsamp-r1: Effective temporal sampling with reinforcement fine-tuning for video llms. In *Adv. Neural Inform. Process. Syst.*, volume 38, pp. 40692–40716, 2025c.

Yunheng Li, Guohong Mu, Hao Li, Shengsheng Qian, Dingwen Zhang, Qibin Hou, and Ming-Ming Cheng. Annotations as rollouts: Efficient and scalable reinforcement learning for video mllms. *arXiv preprint arXiv:2608.20492*, 2026b.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Conf. Empir. Methods Nat. Lang. Process.*, pp. 5971–5984, 2024.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. In *Int. Conf. Learn. Represent.*, pp. 132277–132304, 2026.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Annu. Meet. Assoc. Comput. Linguist.*, pp. 12585–12602, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *Adv. Neural Inform. Process. Syst.*, volume 36, pp. 46534–46594, 2023.

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.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Int. Conf. Comput. Vis.*, pp. 2630–2640, 2019.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Int. Conf. Mach. Learn.*, pp. 28492–28518, 2023.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 18221–18232, 2024.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-SALMONN: Speech-enhanced audio-visual large language models. In *Int. Conf. Mach. Learn.*, pp. 47198–47217, 2024.

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.

Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024a.

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *Int. Conf. Learn. Represent.*, pp. 42055–42079, 2024c.

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.

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. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5288–5296, 2016.

Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. In *Adv. Neural Inform. Process. Syst.*, volume 38, 2025.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *Int. Conf. Learn. Represent.*, pp. 83048–83077, 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.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *Int. Conf. Learn. Represent.*, pp. 98891–98913, 2025.

Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint arXiv:2501.07888*, 2025.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Conf. Empir. Methods Nat. Lang. Process.*, pp. 543–553, 2023.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024a.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024b. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun MA, Ziwei Liu, and Chunyuan Li. LLaVA-video: Video instruction tuning with synthetic data. *Trans. Mach. Learn. Res.*, 2025.

Shanshan Zhao, Xinjie Zhang, Jintao Guo, Jiakui Hu, Lunhao Duan, Minghao Fu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025.

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, volume 32, 2018.

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

Appendix for Attribute-Level Data Self-Evolution for Audiovisual Instruction Tuning

This appendix provides implementation and dataset details, a source-backed comparison of annotation pipelines, benchmark protocols, additional experiments, attribute-coverage analyses, qualitative examples, and the prompt templates used for annotation and evaluation.

CONTENTS

A Limitations **15**

B Implementation Details **15**

C Annotation Datasets **15**

D Comparison of Annotation Pipelines **16**

E Evaluation Benchmarks **16**

F Additional Experiments **17**

F.1 Human Validation Protocol 17

F.2 Effect of Audiovisual Annotation Scope 17

F.3 Evaluation with an Additional Judge 17

F.4 Training Stages and Schedules 18

G Attribute Coverage Analysis **18**

H Visualizations **18**

I Prompts **18**

A LIMITATIONS

First, ASID-1M is constructed via multi-source automatic annotation, ensembling, and verification, and residual noise may remain despite Stage 3 attribute-wise checking, particularly for ASR-dependent speech content and fine-grained timestamps. Second, parts of our evaluation depend on strong proprietary models as judges or downstream solvers (e.g., GPT-4.1/GPT-4o/Seed-1.6/Gemini), which may introduce evaluator bias and makes absolute scores sensitive to the chosen judge. Finally, ASID-1M focuses on short-form videos, and generalization to longer videos or different domains requires further validation.

B IMPLEMENTATION DETAILS

We summarize the stage-wise training configuration of ASID-Captioner in Table 7. All experiments are conducted based on Qwen2.5-Omni and trained via supervised fine-tuning. Across all stages, we use AdamW as the optimizer with a learning rate of 2×10^{-5} and a global batch size of 128. To control computational cost, we enforce a fixed pixel budget per sample, with a reduced total pixel budget in LC. For optimization efficiency and memory scalability, we adopt DeepSpeed with ZeRO-2 in SA and AA, and ZeRO-3 offloading in LC. During SA and AA, training jointly optimizes the projector modules and the language model, while LC adopts LoRA for parameter-efficient adaptation. All models are trained using 32 NVIDIA A800 GPUs with 80 GB memory each, and inference is conducted on a single NVIDIA A800 GPU.

Table 7: Training-stage configuration of ASID-Captioner. SA, AA, and LC denote single-attribute learning, all-attribute learning, and longer-context adaptation (Section 4).

Stage	**SA**	**AA**	**LC**
Purpose	Single-Attribute Alignment	Short-Context Modeling	Long-Context Modeling
Video duration	0–3 min	0–30 s	30 s–3 min
Dataset Items	150k	136k	53k
Total Pixels	20M	20M	12M
Batch size	128	128	128
Optimizer	AdamW	AdamW	AdamW
Learning rate	2e-5	2e-5	2e-5
Warmup Ratio	5%	5%	5%
Precision	Bfloat16	Bfloat16	Bfloat16
Sequence Length	16384	16384	22384
Deepspeed	zero2	zero2	zero3_offload
Train Components	Proj., LLM	Proj., LLM	LoRA

C ANNOTATION DATASETS

LLaVA-Video-178K. LLaVA-Video-178K (Zhang et al., 2025) is a large-scale open-domain video instruction dataset constructed from YouTube, consisting of untrimmed video clips with durations ranging from 5 seconds to 3 minutes. The videos span diverse content categories, including daily activities, instructional videos, cooking, sports, TV shows, and egocentric scenarios, and are curated to emphasize temporal diversity and dynamic scenes rather than static content. From this dataset, we sample approximately 119K videos from the YouTube subset for annotation.

FineVideo. FineVideo (Farré et al., 2024) is a large-scale video dataset consisting of approximately 43K videos with a total duration of about 3.4K hours. The dataset is curated through automatic filtering to retain videos with dynamic visual content and dense spoken language, using criteria such as visual motion intensity and word-level speech density. It emphasizes videos with clear action dynamics and temporally rich audiovisual signals, while excluding static scenes and low-information content. These properties make FineVideo particularly suitable for learning fine-grained temporal alignment between visual events and speech. From this dataset, we sample approximately 16K short-form videos with durations under three minutes for annotation.

D Comparison of Annotation Pipelines

Table 8 compares ASID-Verify with representative pipelines that generate structured captions, correct annotations, or gather evidence iteratively (Chen et al., 2024a; Ju et al., 2024; Xue et al., 2025; Farré et al., 2024; Chen et al., 2026a; Bao et al., 2025; Ma et al., 2026). Structured annotation is not unique to our work, as MiraData, UltraVideo, and FineVideo also organize annotations by aspect or scene. However, prior quality control mainly filters samples or adjusts timestamps at the level of videos, captions, or temporal intervals, without revising erroneous statements or recovering omissions. Iterative pipelines revise other states rather than the caption itself, as Vid-Group/ReCorrect refines interval labels and Omni-Detective accumulates observations before synthesizing the caption. In contrast, ASID-Verify checks individual attributes and carries the repaired caption itself into the next round, and Section 3.4 isolates its effect by comparing attribute-level with caption-level verification.

Table 8: Representative video annotation and correction pipelines, characterized by the operations reported in each source. V and A denote visual and audio input.

Pipeline	Input	Annotation	Check unit	Operations	Iterative state
ShareGPT4Video	V	Integrated caption	Video, caption	Filter, aggregate	None
MiraData	V	Short, dense, 4 aspects	Video	Filter	None
UltraVideo	V	9 aspects, summary	Video-caption pair	Filter, aggregate	None
FineVideo	A+V	Timecoded scenes, QA	Scene boundary	Align, filter	None
AVoCaDO	A+V	Aligned A+V caption	Caption	Filter, aggregate	None
Vid-Group/ReCorrect	V	Query, interval	Query-interval pair	Filter, correct	Interval label
Omni-Detective	A+V	Detailed caption	Observer report	Check, aggregate	Observations
ASID-Verify (Ours)	A+V	Single-/all-attribute	Attribute	Audit, repair	Caption

E Evaluation Benchmarks

video-SALMONN-2. The video-SALMONN-2 test (Tang et al., 2025) contains 483 videos drawn from 14 different domains. Each video lasts between 30 and 60 seconds, with an average duration of 51 seconds. Caption quality is evaluated by an automatic judge that compares generated captions with reference events and categorizes errors into missing, incorrect, and hallucinated events. Incorrect and hallucinated events are treated as hallucination errors, and the final error score is computed by summing the missing rate and the hallucination rate.

UGC-VideoCap. UGC-VideoCap (Wu et al., 2025) consists of 1,000 short-form user-generated videos collected from TikTok. All videos are shorter than 60 seconds and contain at least one meaningful audio segment lasting no less than five seconds. Captions are evaluated by a judge model that assigns scores on a five-point scale across three dimensions, namely visual content, audio content, and descriptive detail. The normalized scores are aggregated into an overall caption quality score.

VDC. The VDC (Chai et al., 2025) benchmark contains 1,027 videos and is designed for evaluating fine-grained visual captioning under a structured protocol. For each video, models generate captions conditioned on five predefined visual aspects, including camera, short summary, background, main object, and detailed description, using aspect-specific prompts. Performance is measured using both accuracy and judge scores for each aspect.

VidCapBench-AE. VidCapBench-AE (Chen et al., 2025) is an automatically evaluable subset of VidCapBench, designed to assess whether video captions provide sufficient and accurate information for controllable text-to-video generation. The benchmark is built on a large collection of short video clips drawn from diverse sources, covering a wide range of visual styles, motions, and scene compositions, with most videos being short-form clips suitable for generation-oriented evaluation. VidCapBench-AE reformulates caption evaluation as aspect-specific question answering, where each video is associated with multiple QA pairs targeting four generation-critical aspects, namely video aesthetics, video content, video motion, and physical laws. For each aspect, caption quality is evaluated using accuracy, precision, coverage, and conciseness, enabling fine-grained assessment of both correctness and completeness with respect to generation-relevant semantics.

Daily-Omni. Daily-Omni (Zhou et al., 2025) is an audiovisual question answering benchmark consisting of 684 videos depicting diverse everyday scenarios collected from multiple platforms. The benchmark includes 1,197 multiple-choice question-answer pairs spanning six task categories. In

our evaluation setting, generated captions are provided as the sole input to a fixed QA model, and performance is measured by question answering accuracy.

WorldSense. WorldSense (Hong et al., 2026) focuses on tightly coupled audiovisual reasoning and contains 1,662 temporally aligned audiovisual clips grouped into eight semantic domains. The dataset includes 3,172 multiple-choice question-answer pairs covering 26 downstream tasks. We evaluate caption quality by measuring how effectively the generated captions support accurate question answering when used as input to a judge model.

Charades-STA. Charades-STA (Gao et al., 2017) is a temporal grounding benchmark built on the Charades dataset, focusing on localizing natural language queries in indoor videos. Each sample pairs a textual query with its corresponding temporal segment within a video. We follow the standard data splits, which include approximately 12.4K training samples and 3.7K validation samples. The benchmark is also widely used to evaluate video LLMs trained for temporal grounding, including recent reinforcement fine-tuning methods (Li et al., 2025c). In contrast, our caption-based evaluation setting performs temporal localization using captions only, and performance is reported using mean IoU and recall at multiple IoU thresholds.

F Additional Experiments

F.1 Human Validation Protocol

We manually inspect 1,500 videos sampled across duration ranges. Two annotators independently label each annotation as Pass, Borderline, or Fail with respect to factual consistency, coverage of major events, hallucination, and temporal consistency, and a third annotator adjudicates disagreements. The overall usability exceeds 98%. Usability is an acceptance criterion rather than a guarantee that every timestamp and fine-grained statement is correct, and it is distinct from the automatic No-Error, No-Missing, and Clean diagnostics.

F.2 Effect of Audiovisual Annotation Scope

Table 9 examines how the scope of annotation affects the automatic diagnostics before attribute-level refinement. Adding timestamps and audio lowers both No-Error and No-Missing, reflecting the difficulty of verifying richer annotations. Because the variants differ in what content is evaluated, this comparison does not imply that audio information is detrimental.

Table 9: Annotation-scope ablation before attribute-level refinement.

Annotation scope	No-Error↑	No-Missing↑
Without audio / timestamps	60.6	80.1
Without audio	40.4	77.5
All	35.2	72.9

F.3 Evaluation with an Additional Judge

To assess sensitivity to the evaluator, we re-evaluate captions with Qwen3.5-27B as the judge. As shown in Table 10, ASID-Captioner obtains lower missing and hallucination rates than AVoCaDO, while Gemini-3-Pro achieves the best scores on both metrics. This comparison complements the main evaluation but does not rule out evaluator bias.

Table 10: Additional caption evaluation using Qwen3.5-27B as judge.

Caption model	Missing↓	Hallucination↓
Gemini-3-Pro	22.0	26.8
AVoCaDO	23.9	31.0
ASID-Captioner	22.3	28.1

F.4 Training Stages and Schedules

Table 11 reports the cumulative checkpoints of the three-stage recipe used in the main experiments. AA substantially reduces missing content at the cost of higher hallucination, and LC further lowers both. To examine whether the stage order matters, Table 12 compares alternative schedules for the SA and AA training signals. These schedules are trained separately from the cumulative checkpoints, so their absolute numbers are not directly comparable to Table 11. Mixed SA/AA training obtains lower missing and hallucination rates than SA→AA, whereas reversing the order trades higher missing for lower hallucination. We therefore regard the staged order as an implementation choice rather than a requirement.

Table 11: Performance of cumulative training checkpoints for captioning.

Checkpoint	**video-SALMONN-2 testset**			**VDC Detailed**
	Miss ↓	Hall. ↓	Total ↓	Acc / Score ↑
After SA	42.1	12.8	54.9	36.1 / 1.6
After SA→AA	24.8	19.9	44.7	40.4 / 1.8
After SA→AA→LC	23.4	18.3	41.7	41.7 / 1.9

Table 12: Alternative schedules for the SA and AA training signals, trained separately from the checkpoints in Table 11.

Schedule	Missing↓	Hallucination↓
SA only	43.3	12.6
SA → AA	26.8	20.1
AA only	27.5	19.9
AA → SA	27.8	19.6
Mixed SA/AA	26.4	19.3

G Attribute Coverage Analysis

Figure 6 presents the hierarchical organization of fine-grained semantic categories used in our analysis. The hierarchy consists of three levels, progressing from coarse to fine granularity. The first two levels are manually defined and serve as the annotation schema during dataset construction, specifying high-level and mid-level semantic dimensions such as scene, action, object, and speech. The third, finest-grained level is automatically derived from the completed attribute-structured captions via semantic grouping. For each second-level category, Seed-1.6 groups the corresponding caption spans into fine-grained categories. GPT-5.2 is then used for post-processing to merge redundant categories and normalize names.

H Visualizations

Figures 7 to 13 visualize captions generated by the ASID-Captioner across diverse real-world video scenarios, including natural landscapes, urban scenes, narrative content, human interactions, and sports footage. The results demonstrate that the model produces fine-grained, temporally coherent, and well-grounded captions under different instruction settings, highlighting its robustness and generality in realistic settings.

I Prompts

We summarize the prompts used throughout the data pipeline and evaluation. For data annotation, we first generate detailed timestamped captions from subtitle-annotated clips (Figure 14) and ensemble multi-source captions into a single coherent narrative (Figure 15). For quality control, we verify and complete captions by identifying important visual and speech deficiencies with structured JSON outputs (Figure 16), followed by attribute-wise auditing to report only verifiable errors, missing

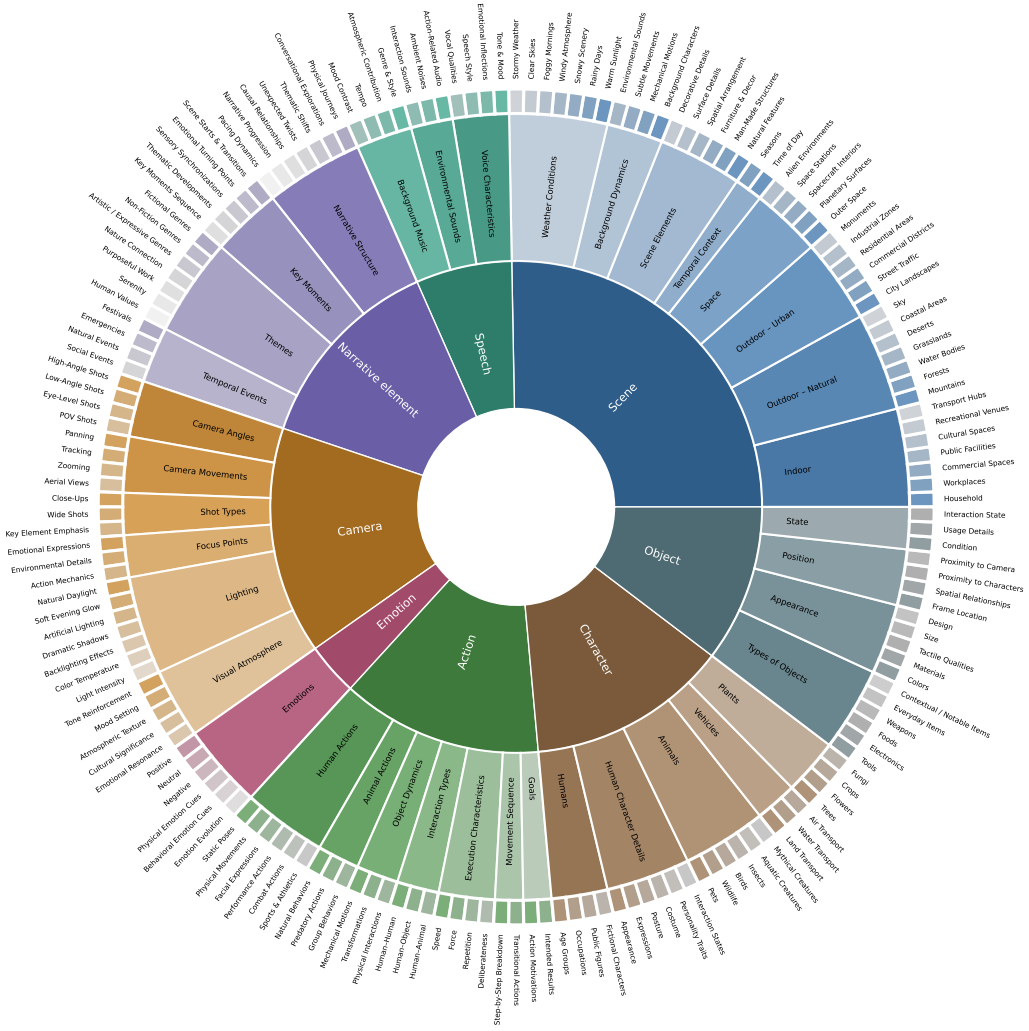


Figure 6: Attribute taxonomy and dataset-wide coverage statistics. The sunburst visualizes the hierarchical attribute taxonomy used in our dataset, where each leaf node corresponds to a fine-grained attribute.



Prompt: Provide a comprehensive description of all the content in the video, leaving out no details, and naturally covering the scene, characters, objects, actions, narrative elements, speech, camera, and emotions in a single coherent account.

Response: At 0s, the video opens with a breathtaking aerial view of a lush mountain valley where mist clings to rolling green fields dotted with quaint red-roofed houses; towering rocky cliffs rise on either side, one crowned by a cascading waterfall that glistens under soft diffused light, while snow-capped peaks peek through distant clouds. The camera slowly pans across the landscape, revealing more of the village nestled among the greenery, then zooms out slightly to show the vastness of the valley and surrounding mountains. A gentle serene instrumental piece-likely played on a piano or similar keyboard instrument, slow-paced and contemplative-fills the audio, creating a calm meditative atmosphere perfect for relaxation. At 10s, white elegant serif text appears over the valley: 'THE ALPS' at the top, below it Chinese characters translating to 'Alps', and at the bottom '4K SCENIC RELAXATION FILM'. The camera continues its smooth pan before transitioning. At 18s, the frame shifts to a sunlit meadow where golden grass stretches toward scattered wooden huts and small rustic cabins, with majestic snow-dusted mountains rising in the distance; warm sunlight bathes the landscape, casting long shadows and highlighting vibrant greens and earthy browns, evoking a peaceful pastoral mood. Text reading 'SCENIC RELAXATION' appears in the bottom left corner, and 'ALPE DI SUISI' is displayed in the bottom right. At 26s, the scene changes to a winding road curving through a verdant valley leading to a charming white church with a tall pointed steeple perched atop a grassy hill; behind it looms a grand snow-capped peak under a bright blue sky with fluffy clouds, and a person walks along the road adding scale to the picturesque setting. Text 'SCENIC RELAXATION' remains in the bottom left, and 'BERCHTESGADEN' is shown in the bottom right. At 34s, an aerial shot reveals Neuschwanstein Castle-a fairytale-like structure with multiple towers and turrets-nestled among autumn-hued trees in shades of orange and yellow; the castle overlooks expansive green fields and a lake, with soft evening light enhancing its fairy-tale allure. Text 'SCENIC RELAXATION' stays in the bottom left, and 'NEUSCHWANSTEIN CASTLE' is visible in the bottom right. At 42s, the frame transitions to Lake Bled's sunset: a small island with a white church featuring two towers sits in the middle of the calm turquoise water, which reflects the golden glow of the setting sun; distant mountains fade into a hazy horizon, and the sun creates a beautiful lens flare effect. Text 'SCENIC RELAXATION' is in the bottom left, and 'LAKE BLED' is in the bottom right. At 50s, the scene shifts to the Church of St. Tomasj-a small white chapel with a red-topped tower-sitting alone on a grassy hill surrounded by bare trees hinting at late autumn or early winter; rolling hills and valleys stretch into the distance under a clear sky, and the camera provides a wide scenic view emphasizing solitude and tranquility. Text 'SCENIC RELAXATION' remains in the bottom left, and 'CHURCH OF SV. TOMAZ' is in the bottom right. At 57s, the final shot shows Grindelwald's valley: a small town nestled between steep rocky mountains-one with a prominent snow patch-under a partly cloudy sky, with lush green forests surrounding the settlement. Text 'SCENIC RELAXATION' is in the bottom left, and 'GRINDELWALD' is in the bottom right. Throughout the video, there is no narration or dialogue-only the soothing instrumental music-allowing viewers to immerse themselves in the tranquil beauty of each location. The camera alternates between sweeping aerial views showcasing landscapes like the valley's mist-shrouded cliffs, Alpe di Suisi's golden meadows, Berchtesgaden's church, Neuschwanstein's autumnal turrets, Lake Bled's reflective waters, St. Tomasj's isolated hilltop, and Grindelwald's forested valley and intimate ground-level shots highlighting details such as the church's steeple, castle's towers, and chapel's red roof, balancing grandeur with intimacy. Lighting varies from soft diffused light in the opening valley to warm golden hues at sunrise/sunset, enhancing moods ranging from mystical to serene to fairy-tale-like. Each transition flows smoothly, weaving together diverse Alpine landscapes to create a cohesive relaxing journey through stunning natural beauty.

Figure 7: Visualization of captions generated by the ASID-Captioner.

items, and expressiveness issues (Figure 17), and audit-guided refinement to minimally fix reported issues while preserving the original timestamp structure (Figure 18). For evaluation, we use caption-grounded multiple-choice QA (Figure 19) and caption-based temporal grounding (Figure 20) to assess factual coverage and temporal localization from textual captions.



Prompt: Provide a comprehensive description of all the content in the video, leaving out no details, and naturally covering the scene, characters, objects, actions, narrative elements, speech, camera, and emotions in a single coherent account.

Response: At 0s, the video opens with a sunlit urban sidewalk lined with parked cars and ornate buildings where a woman with long brown hair wearing a red cardigan over a light top and layered scarf walks briskly beside a man in a dark jacket and blue shirt; the camera follows them from behind initially, then shifts to a side view showing the woman gesturing animatedly as she speaks: "I thought that the dream space would be all about the visuals, but it's more about the feel of it." At 7s, her tone turns curious and intense as she continues: "My question is what happens when you start messing with the physics of it?" The camera cuts to a close-up of the woman looking directly at the man with an earnest expression while he stands still, listening intently against a backdrop of a white van and city street. At 12s, the camera pulls back to show both walking down the narrow street flanked by tall, classic-style buildings, their pace steady as they move forward. At 23s, the sky above the city suddenly warps into a surreal, multi-layered landscape of mirrored, geometric structures twisting the familiar streets below into a disorienting maze; the camera tilts upward to reveal this otherworldly expanse, emphasizing its complexity and scale, while the audio transitions to a deep, resonant, mysterious ambient soundscape filled with low-frequency hums and echoing tones that build suspense and wonder. At 45s, the camera returns to the pair, capturing the woman's wide-eyed amazement and the man's thoughtful gaze as they pause to take in the sight; the woman says: "Something isn't it?" At 56s, the man responds calmly: "Yes, it is," his voice carrying quiet awe. The camera alternates between close-ups of their faces highlighting the woman's wonder and the man's contemplation and wide shots of the warped sky underscoring the surrealism of the moment, then pans across the complex, symmetrical, three-dimensional structure hovering between the buildings to create a dizzying effect before returning to the two standing on the sidewalk, looking up with fascination. The video concludes with a final sweeping shot of the surreal, geometric cityscape, reinforcing the theme of reality bending under unexpected forces.

Prompt: Describe the scene, characters, objects, actions in the video in detail. Write your answer as one coherent paragraph.
Response: A sunlit European-style city street appears under bright daylight, lined with ornate buildings, parked cars (including a black Mercedes), and pedestrians moving along the sidewalk. The setting then shifts into a surreal mirrored cityscape: towering interconnected buildings fold into a vast multi-layered structure made of countless smaller buildings, forming an intricate three-dimensional maze with a dark cavernous space at its center. A woman with long brown hair, wearing a red cardigan over a white top and a light scarf, walks briskly while gesturing expressively. Beside her, a man in a gray jacket over a blue shirt and dark pants follows closely with a serious expression, and later looks up in awe at the transformed architecture.

Prompt: Describe the actions in the video in detail. Write your answer as one coherent paragraph.
Response: The woman gestures with her hands as she walks briskly, and later tilts her head up in awe. The man keeps pace beside her, looking down toward her at first before also lifting his gaze to the sky. Above them, the surreal cityscape appears to rise from the street, twisting and expanding into a gravity-defying mirrored structure, before the viewpoint shifts into a top-down perspective that emphasizes the scale and disorienting geometry of the transformation.

Figure 8: Visualization of captions generated by the ASID-Captioner.



Prompt: Provide a comprehensive description of all the content in the video, leaving out no details, and naturally covering the scene, characters, objects, actions, narrative elements, speech, camera, and emotions in a single coherent account.

Response: At 0s, the video opens with a chaotic crowd of photographers holding cameras, their flashes illuminating a dark night. At 2s, the scene shifts to a close-up of Detective Conan Kudo Shinichi-a high school detective with spiky black hair, wearing a red bow tie and suit-standing against a starry sky dotted with pink laser beams, his expression determined. At 4s, a newspaper headline introduces him: 'His name is Kudo Shinichi,' followed by another headline reading "High School Student Detective Solves Another Case." At 5s, he says, 'I'm a high school detective, Kudo Shinichi.' At 8s, the setting transitions to a lively amusement park at night where Conan and Ran Moriyama, his childhood friend, stand among crowds; Ran points excitedly, and Conan looks on with curiosity. At 9s, Conan says, 'I went to an amusement park with my childhood friend, Moriyama, and witnessed a suspicious transaction between two men in black suits.' At 13s, a close-up shows hands exchanging a purple box filled with money, hinting at a secret transaction. At 15s, the scene moves to a dark, grassy area near a castle-like structure; Conan crouches, fixated on the transaction, but doesn't notice a man in a black coat approaching from behind. At 18s, Conan says, 'I was so obsessed with watching the transaction that I didn't notice another friend approaching me from behind.' At 24s, the man forces Conan to drink poison; Conan lies on the ground, eyes wide with shock, as the man's hand approaches his mouth. At 25s, Conan says, 'I was forced to take poison by that man, and when I woke up, my body had shrunk!' At 30s, Conan wakes up smaller, sitting on the ground looking up with a sad expression as a shadowy figure stands over him. At 32s, the scene shifts to a library where Conan-now in a blue uniform-stands before a stern older man Dr. Agasa, who warns him about being exposed. At 33s, Conan says, 'If they find out that Kudo Shinichi is alive, they will be killed again, and it will also affect the people around them.' At 39s, Dr. Agasa advises Conan to hide his true identity; Conan smiles and gestures with his hands while speaking, then changes into a new outfit. At 40s, Conan says, 'With the help of Dr. Agasa, I decided to hide my true identity.' At 44s, Conan is asked for his name by Ran, and he quickly responds, 'My name is Edogawa Conan.' At 46s, Conan runs towards Ran outside a building labeled 'Mori Detective Agency. At 47s, Conan says, 'In order to get information about them, I ran to Ran's house, where her father is a detective.' The video ends with a shot of the 'Mori Detective Agency' sign on a building facade.

Prompt: Describe the scene, characters, objects, actions in the video in detail. Write your answer as one coherent paragraph.
Response: The video features a bustling night scene with flashing cameras, a starry sky, a bright moon, pink laser-like streaks, and a dark nighttime amusement park with colorful lights, grassy areas, and a castle silhouette. There's also a library filled with bookshelves, a small room with red drawers, and a street at night with a 'Mouri Detective Agency' sign. Characters include Detective Conan—a high school detective with spiky black hair, blue eyes, and a red bow tie—and his childhood friend Ran—a girl with long brown hair wearing a light blue jacket and denim skirt—who later becomes a detective. Other characters include a man in a black coat, a man in a purple suit, an older man with glasses and a white lab coat, and a man in a light green jacket. Objects include cameras, a newspaper, a purple cash register, a briefcase, a syringe, and a small bottle.

Figure 9: Visualization of captions generated by the ASID-Captioner.



Prompt: Provide a comprehensive description of all the content in the video, leaving out no details, and naturally covering the scene, characters, objects, actions, narrative elements, speech, camera, and emotions in a single coherent account.

Response: At 0s, the video opens with a high-angle wide shot capturing six players seated around a black felt poker table in an elegantly decorated indoor room featuring red armchairs, plush seating, wooden floors, and warm lighting; on-screen graphics include GSN logo top right, and text at the bottom. The players include Esfandiari in a light-colored jacket with glasses, Laak in a dark hoodie, Ivey in a blue shirt, Elezra in a black leather jacket, and Negreanu in a gray sweater-all deeply focused on their hands and chips. A male voice says "High stakes poker with six players." At 3s, another voice adds "I wonder who's going broke next," followed by a third saying "Beat him in the pot call." At 9s, a voice states "Phil Ivy has over a million dollars now," then continues "Phil Ivy could actually have a hamburger and have enough money to pay off his million dollar bet with Drawn" while on-screen text identifies Ivey as having over $1 million in chips. At 21s, a voice says "He can actually have a whole barbecue right now if he wants," then another adds "Punish the Unabomber." At 31s, a voice says "I'm easily punishable," followed by "There's many weaknesses in my game that people can exploit." At 37s, a voice says "You can ask Durr about them," then adds "He'll give you the list." At 40s, a voice says "He's very good here," then another says "We might have to pay him something." At 42s, a voice asks "What's it?" then follows with "30 what?" as the camera zooms in for a close-up low-angle shot of a player's hand placing a chip into the pot, with on-screen graphics showing Ivey's name, his cards Ace of Diamonds and 6 of Diamonds, and the pot size increasing from $6, 300 to $14, 100. Throughout the sequence, the camera alternates between close-ups of players' hands, chips, and cards, medium shots of the table, and wider views of the room, emphasizing the strategic tension and high stakes of the game. The atmosphere remains intense yet conversational, with players discussing strategies, chip counts, and potential outcomes-creating a dynamic blend of competitive focus and casual banter.

Prompt: Describe the scene, characters, actions in the video in detail. Write your answer as one coherent paragraph.
Response: A lively indoor poker room features red armchairs, patterned pillows, wooden floors, and warm lighting, where six players sit around a black felt table covered with poker chips, stacks of cash, and a pot counter that rises from 6,300 to 14,100. One player wears a light gray jacket and glasses, while another has his hood up in a dark hoodie; a man in a black leather jacket is also present, alongside players labeled Ivey (holding an Ace and 6 of diamonds later) and Elezra (with an Ace and 7 of clubs). The atmosphere feels tense yet playful as hands move to place chips and handle cards, accompanied by the steady clinking of chips against the felt throughout the sequence.

Prompt: Describe the actions in the video in detail. Write your answer as one coherent paragraph.
Response: Players handle poker chips, place bets, and repeatedly adjust their stacks as the hand progresses. Esfandiari looks down at his cards while Laak methodically stacks and re-stacks chips. Ivey raises, Elezra calls with a smile, and the dealer deals cards as more chips are pushed into the pot. Ivey raises again and the pot grows to $14,100; Negreanu slowly adjusts his chips, and Dwan later raises the pot further to $17,600.

Figure 10: Visualization of captions generated by the ASID-Captioner.



Prompt: Provide a comprehensive description of all the content in the video, leaving out no details, and naturally covering the scene, characters, objects, actions, narrative elements, speech, camera, and emotions in a single coherent account.

Response: At 0s, the video opens with a close-up of two men in a dimly lit taxi at night-one with curly dark hair in a black coat, the other with short light brown hair in a dark jacket over a checkered shirt-both looking forward with serious expressions; the taxi's interior is dark, with blurred city lights visible through the windows. At 1s, the curly-haired man speaks in a calm, analytical tone: 'When I met you for the first time yesterday, I said Afghanistan or Iraq.' At 4s, he continues: 'You looked surprised.' At 5s, the short-haired man asks: 'How did you know?' At 6s, the curly-haired man replies: 'I didn't know, I saw.' The scene shifts to a lab with shelves of glass beakers and equipment, where the short-haired man stands with a rigid posture, looking down with a concerned expression. At 8s, the curly-haired man says: 'Your haircut, the way you hold yourself, says military.' At 10s, he adds: 'But your conversation as you entered the room has been a bit different from my day.' The scene cuts back to the taxi, where the curly-haired man looks out the window with a determined expression. At 13s, he says: 'Said trained at Bart, so army doctor, obvious.' The scene returns to the lab, where the short-haired man looks down with a somber expression. At 17s, the curly-haired man says: 'Your face is tanned.' At 19s, he continues: 'But no tan above the wrists.' At 20s, he adds: 'You've been abroad, but not sunbathing.' A close-up shows the short-haired man's arm, with the curly-haired man's hand holding a phone near it. At 22s, the curly-haired man says: 'Your lip's really bad when you walk.' At 24s, he continues: 'Don't ask for a chair when you stand like you've forgotten about it.' At 27s, he adds: 'So he's partly psychosomatic.' At 29s, he says: 'That says the original circumstances of the injury were traumatic.' At 32s, he continues: 'Wounded in action, then.' At 33s, he adds: 'Wounded in action, sun tan, Afghanistan or Iraq.' The scene shifts back to the taxi, where the curly-haired man looks at the short-haired man with a knowing expression. At 37s, the short-haired man says: 'You said I had a therapist.' At 39s, the curly-haired man replies: 'Of course you've got a therapist.' At 41s, he adds: 'Then there's your brother.' At 43s, he continues: 'Your phone is expensive, email enabled, MP3 player.' At 45s, he says: 'You're looking for a flasher.' At 46s, he adds: 'You wouldn't waste money on this.' At 47s, he continues: 'It's a gift, then.' At 48s, he says: 'Scratches, not one.' At 49s, he adds: 'Many over time.' At 50s, he continues: 'It's been in the same pocket as keys and coins.' At 52s, he says: 'That's something nice people don't treat as one luxury item like this.' At 54s, he adds: 'So it's had a previous owner.' At 56s, he says: 'Next bit's easy, you know it already.' The camera zooms in on the short-haired man's phone, showing an engraving on its back. At 57s, the short-haired man asks: 'The engraving?' The final shot is a close-up of the phone's engraving, which reads 'Harry Watson from Clara.' Throughout the video, a low, tense, and atmospheric electronic score plays in the background, creating a sense of suspense and focus. The camera alternates between close-ups of the men's faces, the short-haired man's arm, and the phone, emphasizing the curly-haired man's meticulous observations and the short-haired man's growing realization. The taxi's dim interior and lab's cluttered shelves create a focused, investigative atmosphere, while the short-haired man's rigid posture and somber expressions highlight his growing realization of his own condition.

Prompt: Describe the objects in the video in detail. Write your answer as one coherent paragraph.
Response: The clip features a taxi, a laboratory lined with shelves of glassware, a beaker and pipette used on a workbench, and a mobile phone with a scratched case. A close-up highlights an engraving on the phone reading “Harry Watson from Clara,” emphasizing the object’s provenance.

Figure 11: Visualization of captions generated by the ASID-Captioner.



Prompt: Provide a comprehensive description of all the content in the video, leaving out no details, and naturally covering the scene, characters, objects, actions, narrative elements, speech, camera, and emotions in a single coherent account.

Response: At 0s, the video opens on a high-angle wide shot of an intense NBA2K ONLINE 2 matchup between the Memphis Grizzlies (MEM, white jerseys) and the Orlando Magic (ORL, blue jerseys) on a polished wooden court labeled "BERLIN GAME 2026" presented by Tissot; the arena is packed with spectators behind barrier boards featuring Revolut logos, and on-screen graphics show the NBA2K ONLINE 2 logo at top left, a bottom scoreboard reading MEM 107–107 ORL with 2:38 remaining in the 4th quarter, a "4K ULTRA HD" watermark at the right, and Chinese text with a bilibili logo at bottom right. A white-jersey ballhandler works near the three-point line under pressure and the play quickly turns physical at the rim; at 5s, the commentator bursts out "Jackson!" and then immediately "No whistle! Turnover!" as possession flips and both teams sprint into transition. At 9s, another call—"Take that with you!"—rides over the roar of the crowd and the squeal of sneakers as the fast break develops, with the camera tracking the bouncing ball and cutting in on determined faces and rapid footwork. Around 15s, a blue-jersey player attacks the basket and collides mid-air with multiple white-jersey defenders, including Jackson (#45) and Williams Jr., sending bodies tumbling to the floor as the commentator reacts, "Here we go the other way," followed by a joking "Gotcha airball," and then, "I'm sorry, Grady Jackson," briefly lightening the tension. Referees in black-and-white striped shirts hover nearby as teammates gasp and look on. At 21s, the commentator exclaims, "Anthony Black got me out of my seat! Did you see that?" and at 24s adds, "Of the year candidate," setting up the highlight moment. At 25s, a new sequence shows a blue-jersey ballhandler driving forcefully through several white-jersey defenders into the paint; at 26s, the commentator shouts, "That is the dunk of the year! I'm calling that right now!" as the player explodes upward and hammers the orange ball through the rim. The broadcast then emphasizes that same dunk via multiple replay angles—cutting between a wider court view, a lower/side perspective that accentuates the takeoff and contact, and a closer look near the basket—before the final frame holds on the dunker landing and looking upward with a triumphant expression while the crowd remains fully engaged in the electric atmosphere.

Prompt: Describe the objects in the video in detail. Write your answer as one coherent paragraph.

Caption: The clip features an orange basketball and a broadcast scoreboard overlay showing "4 BERLIN GAME 2026 Presented by TISSOT," with the score tied at 107–107 and 2:38 remaining in the fourth quarter. The polished wooden court is labeled "BERLIN GAME 2026" and "NBA," and includes sponsor markings such as Tissot, Revolut, and League Pass. A digital game clock displays the remaining time, while on-screen graphics include the "NBA2K ONLINE2" logo at the top left, a "4K ULTRA HD" watermark at the top right, and Chinese characters with a bilibili logo at the bottom right.

Figure 12: Visualization of captions generated by the ASID-Captioner.



At 0s, the video opens on the deck of a grand ship bathed in warm golden sunset light, with white railings framing the scene and a sky painted in orange and pink hues; a gentle melancholic orchestral score plays softly in the background. A woman with curly reddish-brown hair styled in an updo adorned with small earrings and a delicate necklace stands beside a man in a dark coat over a crisp white shirt, their hands clasped gently as they gaze at each other. At 4s, the man speaks in a calm, reassuring tone: "Now close your eyes." The woman responds with a soft smile, her voice filled with trust: "Go on," then closes her eyes slowly, her expression softening into quiet anticipation. At 12s, the man guides her hand to grip the metal railing firmly while saying: "Step up." He continues, his voice steady: "Hold onto the railing." At 18s, he adds: "Keep your eyes closed." The woman whispers: "Don't peek." The man reassures her: "I'm not." At 25s, the man says: "Step up onto the rail." The woman steps carefully toward the edge, her feet hovering just above the deck's surface, while the man holds her arm for balance. At 30s, the man urges: "Hold on." At 31s, he repeats it again. At 32s, he reminds her: "Keep your eyes closed." An overhead shot captures them standing side by side on the deck, arms outstretched like birds ready to take flight. At 34s, the man asks: "Do you trust me?" The woman replies with a confident, loving voice: "I trust you." At 37s, the man lifts her slightly off the ground, holding her securely against him as she spreads her arms wide, embracing the feeling of being lifted. At 46s, the man says: "All right, open your eyes." The woman opens her eyes, her face lighting up with wonder as she looks down at the water below, which shimmers with the sunset's glow. She exclaims joyfully: "I'm flying, Jack!"The man smiles warmly, his eyes crinkling with affection, as they stand together on the deck, arms outstretched, savoring the moment. The camera alternates between close-ups of their faces capturing the woman's joyful grin and the man's tender look, and wide shots of the ship's deck emphasizing the vastness of the ocean and the warmth of the setting sun, enhancing the heartwarming and adventurous mood of their shared experience.

Camera **Sence** **Character** **Emotion** **Narrative element** **Action** **Speech**

Figure 13: Visualization of captions generated by the ASID-Captioner.

Prompt template for generating detailed timestamped captions from subtitle-annotated clips.

You are a video captioning assistant.

You will be given a video clip with (i) spoken dialogue subtitles tagged with speaker names and time ranges, and (ii) separate timestamp cues.

Task: Write ONE comprehensive paragraph that seamlessly integrates **Scene, Characters, Objects, Actions, Narrative (with multiple in-sentence timestamps in chronological order), Speech (as direct quotes), Camera, and Emotions**.

Rules: (1) Use ONLY information supported by the clip/subtitles.
(2) Include key moments with explicit timestamps (in seconds) embedded directly in sentences; reflect scene transitions, turning points, and pacing.
(3) Do NOT mention subtitles, on-screen text, or any flashing/overlay elements.
(4) Do NOT use speaker IDs; indicate speakers naturally with dialogue tags (e.g., "she says").
(5) No lists, brackets/parentheses, or bullet-like formatting—output a single flowing paragraph with rich detail.

Inputs: {subtitle-annotated clip}
Output: {one-paragraph timestamped detailed caption}

Figure 14: Prompt template for generating detailed timestamped captions from subtitle-annotated clips.

Prompt template for ensembling multi-source video captions.

You are a video narrative integrator. You will be given three textual descriptions of the same video from different sources.

Task: Merge them into ONE seamless English paragraph that preserves all specific details, maintains chronological order, and synchronizes visual events with corresponding audio/dialogue.

Timestamp rule: Include only key moments using **single integer-second anchors** in the form "At Xs" (no ranges, decimals, parentheses, or duplicate timestamps), spaced naturally across the narrative.

Constraints: Do not mention source names or generation process; do not invent information; keep Whisper dialogue verbatim; output a single paragraph (no lists or brackets).

Inputs: Description 1: {seed}
Description 2: {avocado}
Description 3: {whisper}
Output: {one-paragraph merged narrative}

Figure 15: Prompt template for ensembling multi-source video captions.

Prompt template for cross-source caption verification and completion.

You are a video caption verification and completion assistant.

Inputs: (1) CURRENT_CAPTION, and (2) SOURCES (Whisper ASR + multi-source captions).

Task: (i) identify *important* missing content, separating purely visual deficiencies from speech deficiencies (speech only from Whisper), and (ii) produce a one-time completed caption by supplementing only supported missing details.

Constraints: no fabrication; ignore trivial details; keep timestamps as integer seconds and follow the timestamp style of CURRENT_CAPTION; output **valid JSON with exactly three keys**:
{`non_speech_deficiency_specific`, `speech_deficiency_specific`, `caption`}.

CURRENT_CAPTION: {current_caption}

SOURCES: {whisper, seed, avocado}

Figure 16: Prompt template for cross-source caption verification and completion.

Prompt template for attribute-wise caption auditing.

You are a caption auditor. You are given the video, optional ASR (truth source for speech meaning only), and a caption to audit.

Task: Report **shortcomings only** (errors, missing, expressiveness), and **audit ONLY the items explicitly listed in each dimension's VERIFY list**; do not evaluate anything outside the VERIFY scope.

Evidence rule: output an error only when a contradiction is clear and verifiable. Missing is reported only for clearly present, important items under VERIFY.

Output: Return **strict JSON only** following the provided schema (no markdown, no explanations). Use: errors as `{"snippet": "...", "why": "..."}`, missing as `{"what": "..." "why": "..."}`, and expressiveness as a list of short strings.

Inputs: {video}, {asr (optional)}, {caption}

Figure 17: Prompt template for attribute-wise caption auditing.

Prompt template for audit-guided caption refinement.

You are a caption refiner. You are given a caption and its audit report.

Task: Produce an improved caption by (i) fixing factual errors, (ii) filling important missing items, and (iii) doing minimal language-only polishing for clarity.

Constraints: Use the audit report as the only edit specification; do not add new content beyond it. Keep the original timestamp blocks and their order unchanged. Timestamps must follow "At Xs," (X is integer or one-decimal), with no ranges.

Output: Return **strict JSON only** with keys `improved_caption` and `enhance_summary` (fixed_errors, filled_missing, dropped_forbidden, timestamp_adjustments).

Inputs: {caption}, {audit_report}

Figure 18: Prompt template for audit-guided caption refinement.

Prompt template for multiple-choice question answering based on textual video captions.

You are a precise QA assistant. Your task is to answer multiple-choice questions based ONLY on the video caption provided.

Do not use any outside knowledge or assumptions—your answer must strictly reflect information from the caption. Always output only the capital letter corresponding to your choice (e.g., A, B, C, D). If the caption does not provide enough information to answer the question, output "N/A" instead.

Here is the video caption: {video caption}

Question: {question}
Choices: {choices}

Figure 19: Prompt template for multiple-choice question answering based on textual video captions.

Prompt template for temporal grounding based on textual video captions.

You are a temporal grounding assistant.

You will be given: (1) a long video caption with multiple timestamp anchors like "At 0s, ... At 5s, ...", (2) an event description (a sentence).

Goal: Infer the most likely continuous time interval (start and end in seconds) when the event happens, using ONLY the caption.

Rules: 1) Always try to output a time interval, even if the evidence is partial. Use best-effort inference. 2) Prefer intervals aligned to existing anchors. If the event is mentioned near "At Ts", choose a range that covers that anchor and the most plausible neighboring anchors. 3) If the event is implied by related actions/objects (synonyms/paraphrases), still infer the interval by matching the closest described segment. 4) Output "N/A" ONLY if the caption provides absolutely no usable clue to localize the event (no matching action/object/context anywhere).

Here is the video caption: {video caption}

Question: {question}
Choices: {choices}

Figure 20: Prompt template for temporal grounding based on textual video captions.