Omni-Captioner: Data Pipeline, Models, and Benchmark for Omni Detailed Perception

Ziyang Ma*, Ruiyang Xu*, Zhenghao Xing*, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, Eng Siong Chng, Xie Chen
*Equal Contribution

ICLR International Conference On Learning Representations

We present Omni-Detective (Agentic Data Pipeline), Omni-Captioner (Model), and Omni-Cloze (Benchmark) for advancing omni detailed perception.

The Challenge: "Co-growth" Dilemma

- **The Problem**: Current Omni Language Models (OLMs) face an inherent trade-off. As descriptions grow longer to capture fine-grained details, the amount of hallucinated content also rises significantly.
- **Observation**: Short captions are "safe but incomplete," while long captions risk injecting ungrounded content.

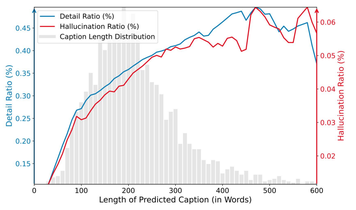

Figure 2: Relationship between caption length, detail coverage, and hallucination on Gemini-2.5-Pro in the detailed captioning task.

Omni-Detective (Agentic Data Pipeline)

- **Concept**: An agentic data generation pipeline that acts like a human detective.
- **How it works:**
  * **Detective Agent**: Orchestrates the process via multi-turn Query-Observation cycles.
  * **Tool Box**: Invokes specialized tools like OCR, ASR, and MLLMs to gather grounded evidence.
  * **Independent Observers**: Analyze raw audio-visual streams to provide precise, cross-checked facts.
- **Benefit**: Decouples detail gain from hallucination growth, yielding high-quality, minimally noisy training data.

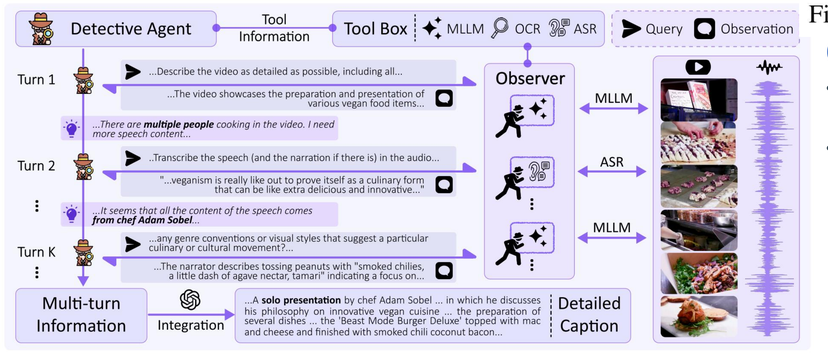

Figure 3: **Omni-Detective.** An agentic data generation and cleaning pipeline integrating specialist tools for omni detailed perception.

Paper | GitHub | Qwen3-Omni-Captioner | Omni-Cloze

Audio/Omni-Captioner (Models)

- **Architecture**: Based on Qwen2.5-Omni-7B backbone.
- **Two-Stage Curriculum Training**:
  - **Stage 1 (Audio Perception)**: Freeze visual encoder to force alignment with sparse but critical audio cues (Audio-Captioner).
  - **Stage 2 (Audio-Visual)**: Jointly optimize both modalities to produce coherent, cross-modal narratives (Omni-Captioner).
- **Impact**: Omni-Captioner captions reach an average of **1,125** words per short video, reflecting extreme detail.

Method	Omni Caption (Truncated)
Qwen2.5-Omni	In this video, we see an intense basketball game in progress during the World Cup Qualifiers. The scoreboard shows that JOR is leading 86–83 … A player from Team JOR dribbles the ball up the court [HALLUCINATION] … Reaching the basket, the player leaps into the air and executes a powerful dunk … [MISSING DETAILS]
Gemini 2.5 Pro	A high-angle, wide shot captures a live basketball game in a packed arena. The scoreboard at the bottom of the screen indicates it's a "WORLD CUP QUALIFIERS" game … The score is PHI 83, JOR 86 … Brownlee elevates over Zaid Abbas (#15) and throws down a thunderous slam … the score updates to PHI 85 – 86 JOR … The final replay shows the ball falling through the net in slow motion ... [MISSING DETAILS]
Omni-Captioner (Ours)	The clip is ... an international basketball game between ... Gilas Pilipinas, and Jordan, played during the FIBA Basketball World Cup ... crowd ambience provides a steady bed of cheers ... The scoreboard shows the score as PHI 83, JOR 86 in the fourth quarter. Jordan wears white jerseys ... while Gilas Pilipinas sports royal-blue ... number 32 from Gilas drives into the lane, receives a pass near the basket ... dribbling past two defenders ... he dunks over a defender wearing jersey number 15 labeled "ABBAS." ... a low-angle slow-motion view that emphasizes his leap and the ball's arc through the net ... The scoreboard now reads PHI 85, JOR 86 with 2:20 left, signaling a narrow lead for the home team ... hashtags such as #FIBAWC. Court-side banners promote "Smart," "TOYOTA," "FOTON TRAVELLER 2.0," "THUNDER 2.0," and "YAMAHA," underscoring the commercial context ... The crowd's diversity—men, women, and children clad in casual attire ranging from T-shirts to hoodies ...

Figure 1: Comparison of the detailed captioning among the Omni-Captioner and other omni models.

Omni-Cloze (Benchmark)

- **The Innovation**: The first cloze-style (fill-in-the-blank) evaluation for detailed perception.
- **Key Features**:
  - Covers **Audio-only, Visual-only, and Audio-Visual** scenarios.
  - **Efficiency**: Reduces LLM calls (1 call per caption vs. 38 for previous benchmarks like VDC).
  - **Reliability**: Includes a **"Not Given"** option to explicitly distinguish between omission and hallucination.
  - **Scale**: 2k video clips with 70k fine-grained cloze blanks across 9 domains.

OMNI-CLOZE Benchmark Visual | Auditory | Audio-Visual

… A sharp electronic [ ] *(A: jingle | B: siren | C: stinger | D: drumroll | E: not given)* with a [ ] *(A: horn | B: clang | C: beep | D: whoosh | E: not given)* punctuates the build as the camera tracks behind the first bowler. The stinger and whoosh signal imminent action just as [ ] *(A: Jakob Butturff | B: Norm Duke | C: Jason Belmonte | D: Kyle Sherman | E: not given)* begins his approach on screen. The left overlay shows "Semifinals - L.A. X Lead [ ] *(A: 0-1 | B: 1-0 | C: 1-1 | D: 2-0 | E: not given)*" with Kingpins [ ] *(A: 96 | B: 126 | C: 117 | D: 146 | E: not given)* …

Figure 4: **Omni-Cloze** utilizes cloze-style MCQ to evaluate models' detailed captioning abilities.

(b) Overall Statistics

Statistics	Number
Total Videos	2320
Total Details	69600
Domain Categories	9
Sub-Categories	47
Avg. Passage Length	444.6 words
Avg. Video Length	34.2 sec

(c) Modality-Wise Statistics

Statistics	Audio	Visual	AV
Total Details	23413	36362	9825
ratio (%)	33.6	52.2	14.1
Avg. Details per Passage	10.1	15.7	4.2

Table 1: Comparison of VDC and our proposed Omni-Cloze benchmark for detailed captioning. Omni-Cloze switches from multi-turn QA to a cloze-style paradigm, enabling a drastic reduction in LLM calls. Omni-Cloze also increases the number of evaluation questions, and expands modality coverage from purely visual to audio-only and audio–visual scenarios.

Benchmark	Evaluation Method	Modality			Questions per Caption	LLM Calls per Caption
		Visual	Audio	AV		
VDC (Chai et al., 2025)	Multiple QA	✔	✗	✗	19	38
Omni-Cloze *(Ours)*	Cloze	✔	✔	✔	**30**	**1**

Table 2: Results on existing benchmarks for evaluating models' detailed captioning ability. The best-performing models in each category are highlighted in **bold**, and the second-best ones are underlined. All the results are obtained from their original papers.

Model	Modality	VDC Detailed		video-SALMONN 2_{test}	
		Acc % (↑)	Score (↑)	Miss % (↓)	Hall % (↓)
Proprietary Models					
GPT-4o (GPT4oTeam, 2024)	V	46.3	2.5	17.0	14.2
Gemini 1.5 Pro (GeminiTeam, 2024)	A + V	43.1	2.2	21.8	16.5
Open-Source Models					
LLaVA-OneVision-7B (Li et al., 2024)	V	41.2	2.1	23.3	27.4
InternVideo2.5-7B (Wang et al., 2025)	V	39.6	2.2	30.8	15.0
Qwen2.5-VL-7B (Bai et al., 2025)	V	44.5	2.4	21.9	17.4
VideoLLaMA3-7B (Zhang et al., 2025)	V	33.4	1.9	44.9	11.6
VideoLLaMA2-7B (Cheng et al., 2024)	A + V	-	-	56.8	**8.9**
video-SALMONN-13B (Sun et al., 2024)	A + V	-	-	52.1	26.6
Qwen2.5-Omni-7B (Xu et al., 2025)	A + V	39.7	2.2	26.3	21.7
video-SALMONN2-7B (Tang et al., 2025)	A + V	46.1	2.5	**10.0**	12.9
Omni-Captioner-7B	A + V	**55.0**	**2.7**	17.8	10.9

Table 3: Results on existing benchmarks using a caption-to-QA cascade evaluation among strong audio-only and omni models.

(a) Audio models.

Model	MMAU	MMAR
Proprietary Models		
GPT-4o Audio	62.4	59.3
Gemini 2.0 Flash	58.6	50.6
Gemini 2.5 Flash	65.6	58.2
Gemini 2.5 Pro	70.0	64.1
Open-Source Models		
SALMONN-13B	58.36	42.5
MiDashengLM-7B	59.4	50.7
Qwen2-Audio-7B	63.3	44.2
Qwen2.5-Omni-7B	65.2	51.8
Audio-Captioner-7B	**70.0**	**59.8**

(b) Omni models.

Model	Video -MME	Video -Holmes	World Sense	Daily -Omni
Proprietary Models				
Gemini 2.0 Flash	64.4	51.6	43.1	60.6
Gemini 2.5 Flash	69.1	52.8	44.6	59.5
Gemini 2.5 Pro	75.0	59.9	53.6	73.6
Open-Source Models				
video-SALMONN-13B	41.8	31.4	26.8	45.0
VideoLLaMA 2-7B	44.4	33.5	26.7	39.9
Qwen2.5-Omni-7B	52.7	35.7	30.6	47.9
video-SALMONN 2-7B	65.9	42.9	44.1	59.7
Omni-Captioner-7B	**67.1**	**48.8**	**48.2**	**67.9**

Table 4: Results on Omni-Cloze among strong audio-only models and omni models.

(a) Audio models.

Model	Acc% ↑
Proprietary Models	
GPT-4o Audio	35.8
Gemini 2.0 Flash	20.0
Gemini 2.5 Flash	42.6
Gemini 2.5 Pro	48.0
Open-Source Models	
SALMONN-13B	10.6
MiDashengLM-7B	19.5
Qwen2-Audio-7B	22.2
Qwen2.5-Omni-7B	25.8
Audio-Captioner-7B	**53.2**

(b) Omni models.

Model	Visual%↑	Audio%↑	AV%↑	Total%↑
Proprietary Models				
Gemini 2.0 Flash	32.3	31.7	40.1	33.2
Gemini 2.5 Flash	24.4	36.2	41.4	33.0
Gemini 2.5 Pro	40.8	44.1	52.8	43.6
Open-Source Models				
video-SALMONN-13B	3.5	2.3	4.6	3.3
VideoLLaMA 2-7B	7.6	4.3	8.7	6.6
Qwen2.5-Omni-7B	18.3	14.1	21.9	16.6
video-SALMONN 2-7B	37.5	40.3	45.0	39.5
Omni-Captioner-7B	**57.0**	**54.5**	**62.1**	**56.4**

(b) Pearson correlation coefficients (r) between human-assigned Elo scores and benchmark-specific automatic evaluation metrics across models. Higher r values indicate stronger agreement between human judgments and the benchmark metric.

Benchmark	Metric	Pearson's r
VDC	VDCscore	0.86
video-SALMONN 2_{test}	LLM-Judge	0.83
Omni-Cloze	Cloze-Acc	0.91