Temporal and Cross-Modal Alignment for Enhanced Audiovisual Video Captioning

Chen Zhao[1,2⋆], Jiajun Ma[1⋆], Qilong Huang[2], Tiehan Fan[1], Hongyu Li[2], Zhuoliang Kang[2], Xiaoming Wei[2], Jian Yang[1], and Ying Tai[1†]

[1] Nanjing University, State Key Laboratory for Novel Software Technology, China
[2] Meituan, China
https://huggingface.co/TCA-Captioner-ECCV-2026

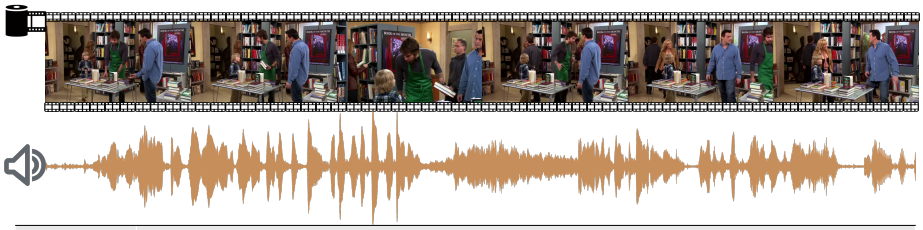
Abstract. While Multimodal Large Language Models (MLLMs) have advanced video understanding, achieving precise temporal and cross-modal alignment in audiovisual video captioning remains a formidable challenge. Most existing approaches suffer from modality detachment and temporal incoherence, failing to accurately bind auditory events to visual entities or capture complex causal dynamics. To address these deficiencies, we propose TCA-Captioner, a framework specifically engineered to enhance Temporal and Cross-Modal Alignment for audiovisual video captioning. We first introduce the Observer-Checker-Corrector (OCC) framework, an iterative refinement strategy that generates high-fidelity, meticulously grounded training data. Leveraging a curated high-density human interaction dataset, TCA-Captioner is optimized to model sophisticated audiovisual interactions. Furthermore, we present TCA-Bench, a diagnostic benchmark utilizing a Decoupled Evaluation Protocol to isolate and quantify model proficiency in audiovisual binding and temporal relational reasoning. Extensive experiments demonstrate that TCA-Captioner sets a new standard for temporally-coherent and synchronized audiovisual narratives.

Keywords: Video Captioning · MLLM · LLM

1 Introduction

Amidst the rapid ascendance of Multimodal Large Language Models (MLLMs) [36, 39, 42, 55, 66, 67, 80, 81, 85], video captioning plays a pivotal role in advancing the frontiers of video understanding and generation [6, 17, 18, 24, 31, 40, 41, 69, 71, 76, 77, 82, 83]. High-quality video captions serve not only as a foundational cornerstone for cross-modal alignment during pre-training but also as a critical mechanism for injecting rich semantic knowledge into downstream tasks [3, 11, 12, 23, 46, 51, 74, 78]. Recent empirical studies suggest that training with detailed and factually grounded captions yields consistent performance gains across a broad spectrum of applications, including video question answering [28], action

⋆ Equal contribution. This work was done while Chen Zhao is an intern at Meituan.
† Corresponding author.



Method	Audiovisual Video Caption
Qwen3-Omni	A young boy in a plaid shirt stands beside a table stacked with books. A man in a green apron, is arranging books on the table. He picks up a book and begins to speak to the boy, who looks up at him with curiosity. The man in the apron then turns and walks away, leaving the boy and another man in a blue shirt standing by the table. The man in the blue shirt gestures with his hands as he speaks, seemingly explaining something to the boy. In the background, a woman in a brown jacket walks past, browsing the shelves. The man in the apron says, "Excuse me, sir, do you work here?" The man in the blue shirt replies, "No, I'm wearing this green apron just for fun. Good luck reading the book, you dummy." The man in the apron then says, "It's just a lot of responsibility, you know." The man in the blue shirt responds, "Hey, did you just yell at my son?" The man in the apron apologizes, saying, "I'm sorry."
Gemini-3-Pro	A young boy of about seven or eight years old approaches a staff member in a bookstore who is wearing a green apron and politely asks, "Excuse me sir, do you work here?" The clerk appears extremely impatient and sarcastic; he snaps around, spreading his arms to gesture at his apron, and yells, "No, I'm wearing this green apron just for fun. Good luck reading a book, you dummy!" before storming off. A man standing nearby witnesses the entire exchange with a look of shock and confusion, muttering to himself, "It's just a lot of responsibility, you know?", clearly startled by the clerk's sudden outburst. Immediately after, the boy's mother walks over and, with a serious and angry expression, demands of the departing clerk (or the man standing there) , "Hey, did you just yell at my son?" The man stammers an apology, his voice sounding nervous as he says, "Uh, I'm sorry if my " before the video cuts off abruptly.
AVoCaDO	A young boy with shaggy blonde hair, stands at a glass display table filled with stacks of books. To his right, a man in a green apron over a dark shirt looks down at the books, holding one. To the right of the man in the apron, a man in a blue button-down shirt and jeans gestures with his hand while speaking. A young boy speaks first, "Excuse me, sir, do you work here?" The man in the apron looks down at the boy with a dismissive expression, still holding the book. The man in the blue shirt stands beside him, looking on. An adult male responds with a loud, aggressive, and condescending voice. "No, I'm wearing this green apron just for fun" he retorts sharply. His voice drips with mockery as he continues, "Good luck reading a book, you dummy!" A loud, comedic "thud" sound effect punctuates his insult, followed by the sound of an audience laughing. Another man, presumably the father, speaks in a calm but concerned tone, "It's just a lot of responsibility, you know?" The man in the blue shirt as he gestures emphatically with both hands while speaking to the woman and the boy. The woman looks at him, her expression shifting from concern to anger. The man in the apron is now out of the frame. The first man's voice returns, now louder and more confrontational, as he yells, "Hey, did you just yell at my son?" The father begins to reply in a placating, slightly apologetic tone, "I'm sorry, my..." before the audio cuts off.
Ours	A young boy with blonde hair, wearing a blue-and-green plaid shirt, stands looking up at a man behind the display table. The man behind the table wears a grey long-sleeved shirt with a bright green apron and holds a stack of books. To the right, another man in a denim button-down shirt stands observing. A woman in a brown suede-textured jacket stands near the bookshelves in the background. The ambient sound is a low room tone until the boy speaks in a polite, high-pitched voice, asking, "Excuse me, sir, do you work here?" The employee in the green apron reacts immediately with hostility. Adjusting the stack of books in his hands, he looks down at the boy with a scowl. His tone is loud and sarcastic as he retorts, "No, I'm wearing this green apron just for fun." Leaning in sharply toward the child to emphasize the insult, he adds, "Good luck reading a book, ya dummy!" The camera cuts to a close-up of the boy, who stands frozen with a neutral, slightly stunned expression. A studio audience laugh track swells, overlapping the end of the employee's sentence. The employee then turns and walks away to the right, moving behind the man in the blue shirt. As he passes, the employee gestures with his right hand—palm up and fingers curled—and complains in a weary tone, "It's just a lot of responsibility, you know?" The man in the blue shirt watches the employee leave with a bewildered expression, his eyebrows raised. Immediately, the woman in the brown jacket, who was previously in the background, walks briskly into the foreground, fixing her gaze on the man in the blue shirt. Her voice is sharp and accusatory as she demands, "Hey, did you just yell at my son?" The man in the blue shirt reacts with panic, his eyes widening, placating gesture and stammers, "I'm sorry if my f—" just as the video cuts off abruptly."

Fig. 1: The challenges for audiovisual video captioning. Existing state-of-the-art models, including Qwen3-Omni, Gemini-3-Pro, and AVoCaDO, still struggle with two fundamental challenges: **Audio-Visual Cross-Modal Binding** and **Audio-Visual Temporal Coherence**. In the provided examples, text highlighted in red indicates errors in Cross-Modal Binding. Text marked with purple underlines identifies failures in Temporal Coherence. In contrast, our model, shown in the bottom row, accurately captures these complex interactions, with correctly grounded and synchronized descriptions highlighted in green. The original video file will be provided in the supplementary material to facilitate a more detailed comparison and observation.

recognition [57], and generative video synthesis [23,44]. Therefore, enhancing the capabilities of video captioning models has emerged as a fundamental pathway toward building more robust video understanding and generation systems.

Despite the remarkable progress of MLLMs, early research in video captioning has remained predominantly vision-centric [6, 18, 63], often overlooking the profound semantic cues embedded within auditory signals. In practice, auditory elements such as dialogues, voiceovers, and ambient soundscapes are indispensable for achieving a holistic and embodied understanding of video content. The emergence of omni-modal MLLMs, such as Qwen3-Omni [60] and Gemini 3 Pro [22], has signaled a paradigm shift toward joint audio-visual reasoning. Concurrently, specialized audiovisual captioners, including AVoCaDO [7], video-SALMONN 2 [49], and UGC-VideoCaptioner [58], have begun to explore the complementary nature of auditory and visual streams to produce more comprehensive narratives.

Nevertheless, two critical challenges in joint audio-visual video captioning remain insufficiently addressed. The first is **Audio-Visual Cross-Modal Binding**, which refers to the precise association between specific auditory events

and their corresponding visual entities. For instance, models frequently struggle to correctly attribute a segment of speech to its specific speaker in crowded scenes or link a transient impact sound to its physical origin, leading to attribute misattribution. The second challenge is **Audio-Visual Temporal Coherence**, which concerns the chronological ordering and causal dependencies across audio-visual modalities. Existing models often fail to capture the fine-grained temporal dynamics between audio signals and visual actions, such as the "sound-before-sight" anticipation or "action-followed-by-sound" feedback, thereby struggling to model the evolving logic of multimodal event streams. As illustrated in Fig. 1, Qwen3-Omni [60] exhibits a pronounced modality isolation in its descriptions; it tends to narrate visual content and auditory elements in a decoupled, sequential manner (i.e., describing all visual scenes before addressing any audio), which entirely neglects the temporal coherence between the two streams. Furthermore, it suffers from significant errors in cross-modal binding. Similarly, both Gemini-3-Pro [22] and AVoCaDO [7] demonstrate localized failures in binding and temporal synchronization within their generated captions. These deficiencies represent a significant barrier to achieving a truly unified perception of synchronized audio-visual content.

To bridge these gaps, we present TCA-Captioner along with a diagnostic benchmark, TCA-Bench, specifically engineered to enhance **T**emporal and **C**ross-modal **A**lignment in audiovisual video captioning. To facilitate high-fidelity model training, we first design a rigorous data synthesis pipeline termed Observer-Checker-Corrector (OCC), powered by the state-of-the-art Gemini 3 Pro. Unlike conventional single-pass annotation, the OCC pipeline adopts an iterative refinement strategy. In this framework, the Observer first generates initial descriptions, which are then rigorously vetted by the Checker to identify unimodal factual hallucinations and cross-modal alignment inconsistencies. By ensuring the narrative is faithful to both individual streams and their joint dynamics, this verification step allows the Corrector to precisely rectify misalignments, thereby producing meticulously grounded captions. Furthermore, to specifically address the intricacies of temporal and cross-modal dependencies, we curate a high-density human interaction Dataset. This dataset is characterized by dense human speech interleaved with complex physical actions, providing a rich corpus of synchronized audiovisual cues that are often absent in existing datasets. Leveraging this high-fidelity data, we optimize TCA-Captioner through a robust post-training regime centered on Supervised Fine-Tuning (SFT). Extensive experiments demonstrate that our TCA-Captioner achieves superior performance in modeling sophisticated cross-modal interactions, effectively setting a new standard for temporally-coherent audiovisual narratives.

To facilitate a more granular assessment of audiovisual video captioning, we introduce TCA-Bench, a diagnostic benchmark that advocates a paradigm shift from static detail recall to the evaluation of dynamic audio-visual entanglement. Unlike existing benchmarks that rely on holistic scoring or "bag-of-events" matching, which frequently neglect the structural coherence of narratives, TCA-Bench specifically addresses the prevalent failure mode of modality detachment.

We provide structured annotations, including Audio-Visual Binding Lists and Temporal Relational Lists, to formalize the fine-grained dependencies between acoustic events and their visual origins. Furthermore, we propose a Decoupled Evaluation Protocol that disentangles foundational uni-modal perception from higher-order relational reasoning. This protocol employs an LLM-based judge to perform targeted, hallucination-resistant verification of cross-modal source localization and chronological ordering. By isolating these synchronization capabilities, TCA-Bench offers a highly interpretable and granular assessment, uncovering the primary bottlenecks of current models in achieving unified and temporally-coherent multimodal understanding. Our contributions can be summarized as follows:

- We introduce the Observer-Checker-Corrector (OCC) framework, which employs an iterative refinement strategy to disentangle multifaceted audiovisual caption generation, guaranteeing the production of meticulously grounded and high-fidelity video descriptions.
- We develop TCA-Captioner, specifically optimized for fine-grained cross-modal binding and temporal coherence, supported by a newly curated dataset of high-density human interactions.
- We propose TCA-Bench, a diagnostic benchmark utilizing a Decoupled Evaluation Protocol to quantify the model's proficiency in audio-visual binding and temporal relational reasoning.

2 Related Works

2.1 Multimodal Large Language Models

The landscape of Multimodal Large Language Models (MLLMs) has undergone a rapid paradigm shift, expanding from basic image-text alignment to sophisticated, all-encompassing sensory perception [56]. Early endeavors, such as LLaVA [35] and BLIP-2 [29], focused on bridging the modality gap by aligning frozen visual encoders with LLMs via specialized adapters, a foundation later scaled by InternVL [8] and VILA [33]. This evolved into temporal modeling with Video-LLaVA [32], ShareGPT4Video [6], and LLaVA-Video [73], which utilized high-quality synthetic data to capture dynamic transitions in silent videos. Concurrently, dedicated models like SALMONN [50] and Qwen2-Audio [13] emerged to tackle non-visual acoustic reasoning and spatial audio understanding [66]. Building upon these modular successes, the field is converging toward Omni-modal MLLMs [22,60], aiming for a holistic interpretation of real-world scenarios by processing interleaved text, vision, and audio tokens within a unified architecture. Industry leaders like GPT-4o [25] and the Gemini series (including Gemini 3 Pro) [22] have demonstrated the efficacy of native multimodal integration, exhibiting superior reasoning in long-context and real-time environments. In the open-source community, the Qwen-Omni series [59, 60] has mirrored this trend: while Qwen2.5-Omni [59] achieved seamless audio-visual streaming, the state-of-the-art Qwen3-Omni [60] further pushes these boundaries by introducing a

"multimodal thinking mode" and significantly reducing latency for extended inputs. Our work is positioned at this frontier, aiming to systematically evaluate and enhance MLLM capabilities in these complex, omnimodal scenarios.

2.2 MLLM for Video Captioning

The integration of Large Language Models (LLMs) has fundamentally reshaped video captioning, evolving from template-based outputs to context-aware narratives. Early VideoLLMs [2, 4, 5, 15, 16, 26, 27, 30, 37, 65, 68, 70] typically coupled pre-trained visual encoders with language backbones to capture dynamic transitions. To enhance granularity, recent models like OwlCap [79] and the Tarsier series [54] leveraged large-scale SFT datasets to balance high-level motion semantics with fine-grained static details. However, these vision-centric models remain "deaf" to the audio modality, limiting their interpretation of real-world videos where acoustic cues are indispensable. To overcome this, the field has pivoted toward omnimodal architectures [20, 38, 45, 48, 60, 64, 72, 75]. Proprietary models like Gemini 3 Pro [22] and open-source breakthroughs such as Qwen3-Omni [60]] treat vision and audio as synchronized streams within a unified transformer space. Despite their general reasoning prowess, they are not explicitly optimized for the dense, temporally-aligned captioning essential for high-fidelity video understanding. While concurrent efforts like Video-SALMONN-2 [49] and UGC-VideoCaptioner [58] employ intensive DPO or focus on specific content types, achieving precise temporal alignment between transient acoustic events and visual frames remains an open challenge. In this work, we aim to enhance temporal and cross-modal alignment for joint audiovisual video captioning, ensuring that visual dynamics and acoustic semantics are harmoniously integrated into high-quality descriptions.

3 TCA-Captioner

3.1 Challenge for Audiovisual Video Captioning

Traditional video captioning models primarily follow a vision-centric paradigm, treat-ing audio as a redundant or auxiliary signal [1, 6, 9, 10, 18, 34, 84]. However, real-world audiovisual content is defined by high-density interactions where semantics are deeply entangled across modalities. Despite recent progress in joint audiovisual captioning, two primary bottlenecks remain largely overlooked: (1) **Audio-Visual Cross-Modal Binding**, the challenge of accurately associating a transient acoustic event (e.g., a specific utterance or a physical impact) with its corresponding visual source in a complex scene; and (2) **Audio-Visual Temporal Coherence**, the requirement to preserve the precise chronological and causal flow between auditory and visual streams. The neglect of these factors leads to "modality detachment," where captions may contain correct individual elements but fail to capture their synchronized relational logic.

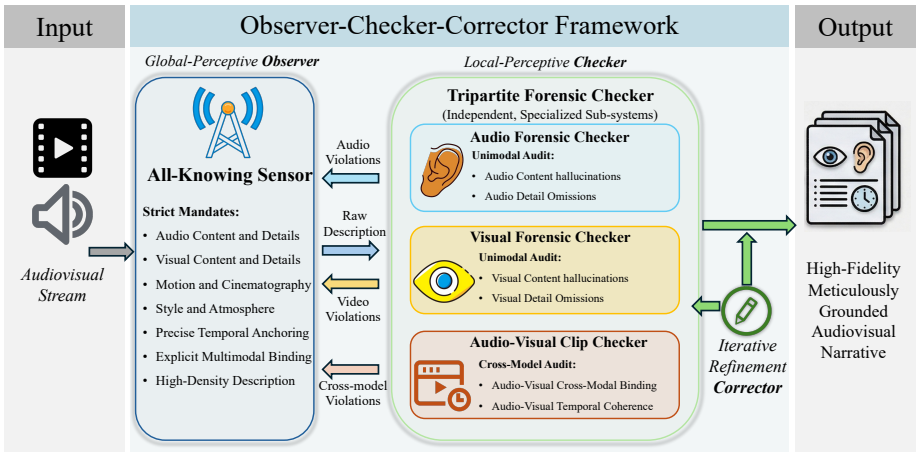


Fig. 2: The proposed Observer-Checker-Corrector (OCC) framework, an iterative data synthesis pipeline designed to generate high-quality audiovisual video captions. The framework consists of three collaborative modules: a Global-Perceptive Observer for comprehensive understanding, a Local-Perceptive Checker for detecting unimodal factual inaccuracies and cross-modal misalignments, and an Iterative Refinement Corrector for precision error correction and narrative enhancement.

3.2 Observer-Checker-Corrector Framework

To address the aforementioned challenges, we propose the Observer-Checker-Corrector (OCC) framework, an iterative data synthesis pipeline designed to generate high-fidelity, meticulously grounded audiovisual narratives. Unlike previous single-pass annotation, OCC employs a "divide-and-conquer" strategy through three collaborative modules, as shown in Fig. 2.

Global-Perceptive Observer The Observer serves as the primary perception engine, functioning as an all-knowing sensor that performs a global analysis of the audiovisual stream. Its objective is to produce a raw, high-density description.The Observer is constrained by a strict mandate: (1) Audio Content and Details: Identify all sound events (speech, foley, ambient noise, background music) and their physical characteristics (loudness, frequency modulation, and timbre). (2) Visual Content and Details: Describe the scene, entities (appearance, actions), objects, lighting, and color palettes. (3) Motion and Cinematography: Catalog all subject movements (e.g., trajectories, gestures) and cinematographic techniques, including camera motion (pan, tilt, zoom) and shot transitions (cuts, fades). (4) Aesthetic Style and Atmosphere: Characterize the overall stylistic tone (e.g., cinematic, documentary, handheld) and the auditory atmosphere (e.g., tense, reverberant, industrial). (5) Precise Temporal Anchoring: All events must be localized with sub-second timestamps (e.g., [01.2s - 01.5s]) to ensure strict alignment with the raw signal. (6) Explicit Multimodal Binding: Avoiding vague pronouns in favor of entity-specific attribution (e.g., grounding a specific segment of speech to its corresponding speaker among multiple candidates in a scene). (7)

High-Density Description: Prioritize a comprehensive inventory of atomic facts over abstract summaries. We employ Gemini-3-Pro, a formidable MLLM, as the Observer to navigate the inherent intricacies of global audiovisual analysis.

Local-Perceptive Checker To guarantee factual integrity, we introduce a Tripartite Forensic Checker consisting of three independent, specialized subsystems. This decoupled verification ensures that unimodal errors do not propagate into cross-modal misalignments.

- **Audio Forensic Checker**: This module performs an independent audit of the audio track to verify auditory hallucinations and identify omissions in granular sound details. Focusing on Verbal Fidelity and Linguistic Integrity, it strictly penalizes phonetic discrepancies and flags "phantom sounds" that lack an acoustic basis. Beyond content validation, the checker scrutinizes fine-grained attributes, such as timbre characteristics, ambient interference, and rhythmic prosody, while rigorously ensuring the temporal coherence of acoustic events to maintain precise onset and offset alignment. We employ Gemini-3-Pro as the audio checker to ensure the factual integrity and perceptual accuracy of the auditory information.
- **Visual Forensic Checker**: Acting on silent video tracks, this module is engineered to audit visual hallucinations and recover omitted fine-grained details. It prioritizes the validation of entity dynamics and spatio-temporal logic, meticulously scrutinizing micro-expressions and motion trajectories to ensure the narrative aligns with the physical causality of the visual scene. Furthermore, it supplements any overlooked granular nuances while rigorously enforcing the temporal coherence of visual events, ensuring that the chronological flow of actions remains logically sound and consistent. We observe that Gemini-3-Pro introduces a significant amount of hallucinations when processing silent video streams. Therefore, we employ the powerful silent video understanding model, Doubao-1.8, as the Visual Checker, which possesses more accurate and fine-grained visual understanding capabilities.
- **Audio-Visual Clip Checker**: This is the core module for evaluating audio-visual cross-modal binding and temporal coherence. It distinguishes between diegetic sounds (on-screen), off-screen sounds, and non-diegetic voice-overs. It specifically targets "lip-sync" discrepancies and ensures that the narrative preserves the chronological sequence of cross-modal interactions. To achieve this, the module employs a sliding-window strategy, partitioning the video into 2-second clips that are processed iteratively to generate detailed verification logs. Notably, each input segment is augmented with its sequence index and corresponding timestamps, facilitating superior calibration of temporal coherence. Given the intensive computational demands of this granular analysis, we utilize Gemini-3-Flash as the underlying engine for the Clip Checker, balancing high-fidelity reasoning with operational efficiency.

Iterative Refinement Corrector The Corrector acts as the final synthesis agent. It receives the initial description from the Observer and the comprehensive audit reports (violation lists) from the tripartite Checkers. Its task is to perform an optimal fusion: rectifying hallucinations identified by the unimodal checkers

Table 1: Token Consumption per 10s Video.

Observer	A-Checker	V-Checker	A/V-Clip-Checker	Corrector	Total
3,500	1,535	3,874	8,705	4,335	**21,949**

and resolving alignment conflicts flagged by the audiovisual checker. Through this iterative feedback loop, the Corrector produces a final caption that is both semantically rich and structurally aligned with the raw sensor data. We employ Gemini-3-Pro as the refinement corrector.

Token Cost As detailed in Table 1, processing a 10-second video consumes approximately 22,000 tokens (both input and output) across the OCC pipeline. The A/V-Checker accounts for the majority of this cost due to its iterative processing of the entire audio-visual sequence.

3.3 High-Density Human Interaction Dataset

Existing video datasets [14, 19, 47] often feature sparse audiovisual events, rendering them insufficient for training models on complex alignment tasks. To address this gap, we curate the High-Density Human Interaction (HDI) Dataset, comprising 3,500 meticulously selected clips sourced from YouTube and cinematic content. The HDI dataset is specifically filtered for "hard cases," such as multi-speaker dialogues in noisy environments, rapid physical actions coupled with impact sounds, and subtle cross-modal causal chains. Each clip is processed through our full OCC pipeline, yielding high-fidelity annotations that provide the necessary supervision for the TCA-Captioner to transition from coarse scene-level understanding to fine-grained, synchronized audiovisual perception. Furthermore, to enhance the diversity of the training distribution, we incorporate an additional 5,000 samples from FineVideo [19] and 7,000 samples from TikTok-10M [14]. Given the extended duration and relatively straightforward content of these videos, we streamlined the annotation process by bypassing the Clip Checker module within the OCC framework. Leveraging this comprehensive multi-source corpus, we perform LoRA fine-tuning on the Qwen3-Omni [60] backbone to develop our final TCA-Captioner.

4 TCA-Bench

4.1 Motivation

While recent audiovisual captioning benchmarks have advanced the measurement of detailed perception, they fundamentally assess only the *static presence* of individual details, not the dynamic interplay between modalities. Video-SALMONN-2 [49] decomposes videos into flat atomic events to compute missing and hallucination rates; UGC-VideoCap [58] assigns a single holistic LLM score

Table 2: Comparison of audiovisual video captioning benchmarks. The upper block compares evaluation capabilities; the lower block compares dataset properties. ✓ = explicitly supported; × = not supported; (P) = partial.

Dimension	video-SALMONN-2	UGC-VideoCap	Omni-Cloze	**TCA-Bench**
A-V Binding Evaluation	×	×	×	✓
A-V Temporal Evaluation	×	×	×	✓
Evaluation Paradigm	Event Matching	Holistic Scoring	Cloze-style MC	Decoupled Checklist
Independent Uni-modal Scoring	×	×	(P)	✓
Structured Cross-Modal GT	×	×	×	✓
Number of Videos	483	1,000	2,340	459
A-V Interaction Density	Low	Low	Low	High

across entangled dimensions; Omni-Cloze [43] transforms evaluation into cloze-style multiple-choice questions to penalize hallucinations on isolated attributes. A common structural limitation underlies all three: they treat multimodal content as an unordered bag of independent facts, ignoring the relational structure that connects modalities. Consequently, none of these paradigms can systematically evaluate whether a model correctly *binds* a sound to its visual source or preserves the *temporal ordering* of cross-modal events, two capabilities we identify as the primary failure modes of state-of-the-art models. Table 2 contrasts these benchmarks with TCA-Bench. To bridge this gap, TCA-Bench introduces a **Decoupled Evaluation Protocol**, specifically designed to disentangle foundational unimodal perception from higher-order relational reasoning.

4.2 Benchmark Construction

Data curation. TCA-Bench targets videos with dense audio-visual interplay, specifically multi-person dialogues coupled with complex physical actions (e.g., variety shows, group interactions)—sourced from YouTube. Such scenes produce frequent, entangled cross-modal events that stress-test audiovisual binding and temporal relational reasoning. We focus on short clips (5–30 seconds), as these densely interleaved segments already pose substantial challenges for current models. As illustrated in Fig. 3, the construction pipeline proceeds in two macro-stages: (I) automated clip curation via multi-modal temporal analysis, and (II) structured annotation on a purpose-built labeler platform.

In Stage I, two parallel analysis streams operate on each source video: a visual stream uses YOLOv8 for person detection to produce a per-frame person-count timeline, while an auditory stream applies Silero VAD to generate speech-activity intervals. The two timelines are fused under a joint co-occurrence constraint, retaining only intervals where multiple persons are visually present *and* speech is simultaneously active, followed by quality filtering on person-detection ratio and speech overlap. Human annotators then perform a final curation pass to ensure sufficient density and diversity of cross-modal interactions. In Stage II, each curated clip enters a purpose-built web-based labeler platform where all three annotation layers are produced through a unified *LLM-draft-then-human-*

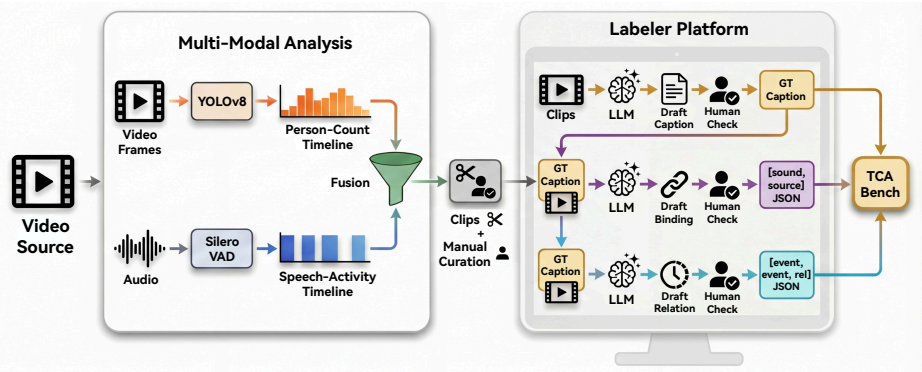


Fig. 3: TCA-Bench data curation and annotation pipeline. *Stage I*: source videos undergo parallel visual (YOLOv8 [53]) and auditory (Silero VAD [52]) analysis; timelines are fused under a joint co-occurrence constraint to extract segments. *Stage II*: each clip receives three ground-truth layers via LLM draft → human validation.

validation paradigm: a multimodal LLM first generates an initial dense caption, which human annotators correct and finalize; the verified caption then serves as context for the LLM to draft binding pairs and temporal triples, each again undergoing human validation.

Structured annotation. Unlike prior benchmarks that rely solely on flat textual ground truths, TCA-Bench formalizes cross-modal dependencies with three complementary annotation layers:

1. **Ground-truth captions** that exhaustively describe visual details (subjects, actions, scenes) and auditory elements (speech, sounds).
2. **Audio-Visual Binding Lists** — structured JSON lists of `[sound, source]` pairs linking each acoustic event to categories: *foreground character*, *off-screen character*, or *environment*.
3. **Temporal Relational Lists** — triples of `[first_event, second_event, relation]` encoding relation types: `A_then_V`, `V_then_A`, or `AV_simultaneous`.

These three layers are explicitly aligned with the evaluation protocol: ground-truth captions support Base Perception scoring, while Binding Lists and Temporal Relational Lists serve as structured references for the cross-modal Binding Accuracy and Temporal F_1 metrics, respectively. Fig. 4 illustrates the three annotation layers on a representative benchmark clip.

4.3 Decoupled Evaluation Protocol

Existing evaluation paradigms conflate uni-modal perception errors with cross-modal reasoning failures, making it impossible to diagnose *where* a model fails. TCA-Bench addresses this with a two-tier **Decoupled Evaluation Protocol** using an independent LLM judge.

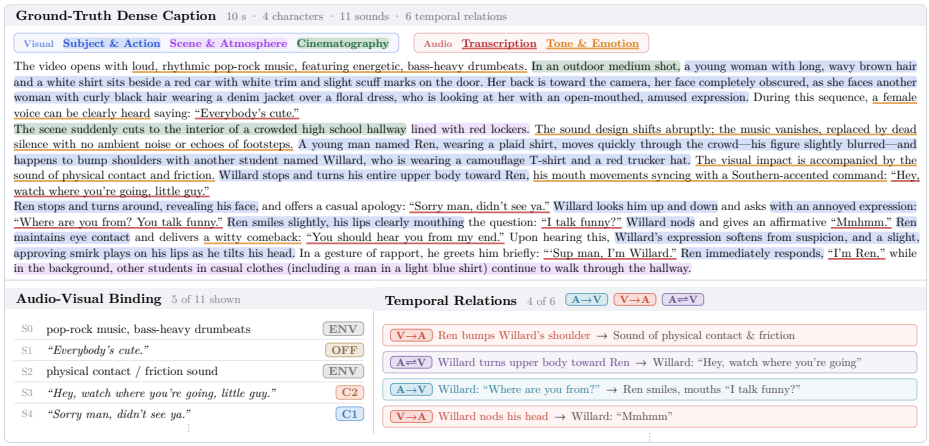


Fig. 4: TCA-Bench ground-truth annotation showcase. **Top:** a dense caption with inline highlights delineating five evaluation sub-dimensions grouped into two modality tracks (Visual: Subject & Action, Scene & Atmosphere, Cinematography; Audio: Transcription Accuracy, Tone & Emotion). **Bottom-left:** audio-visual binding pairs mapping each acoustic event to its source—foreground character, off-screen character, or environment. **Bottom-right:** temporal relational triples encoding cross-modal sequential ($A \rightarrow V$, $V \rightarrow A$) and simultaneous ($A \rightleftharpoons V$) orderings. The clip is chosen for balanced coverage across all sub-dimensions, source categories, and relation types.

Base Perception (Uni-modal). An LLM judge compares the candidate against the GT caption on a 0–10 scale:

- **Audio:** *Transcription Accuracy*—verbatim fidelity of speech content including meaning-critical words and fillers; *Tone & Emotion*—accuracy of the described emotional delivery style and non-speech audio such as music mood and ambient sounds.
- **Visual:** *Subject & Action*—correctness of primary subjects' appearance, quantity, spatial relationships, and sequential movements; *Scene & Atmosphere*—fidelity of the setting, background layout, spatiotemporal attributes, and visual mood; *Cinematography*—accuracy of shot size, camera angles, and camera movement patterns.

Audiovisual Integration (Cross-Modal). This track uses structured GT annotations for targeted checklist evaluation:

- **Binding Accuracy:** For each GT binding pair, the judge performs a two-stage verification: first confirming that the acoustic event is *present* in the candidate caption, then checking whether it is attributed to the *correct source*. A sound is scored as *correct* only when both stages pass. Crucially, this decoupled design ensures that transcription errors (e.g., mishearing speech content) do not penalize binding scores—such errors are isolated in the Base Perception tier. We report accuracy separately for *Character* (foreground + off-screen) and *Environment* sounds.

Table 3: Caption evaluation on the public benchmarks. "A" and "V" refer to the audio and visual modalities, respectively. The results presented above are reproduced using the official code. We strictly adhere to the evaluation protocol established by AVoCaDO to ensure a fair and consistent comparison.

Model	Size	Modality	video-SALMONN-2			UGC-VideoCap			
			Miss ↓	Hall. ↓	Total ↓	Audio ↑	Visual ↑	Detail ↑	Avg. ↑
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
Gemini-3.0-Pro	-	A + V	18.9	14.8	33.7	71.3	76.7	77.8	75.3
HumanOmniV2	7B	A + V	49.2	**12.3**	61.6	45.6	66.3	59.5	57.1
ARC-Hunyuan-Video	7B	A + V	45.7	12.5	58.2	52.7	56.0	55.8	54.8
MiniCPM-o-2.6	8B	A + V	42.2	14.3	56.5	38.6	68.5	57.7	54.9
Qwen2.5-Omni	7B	A + V	41.7	15.4	57.1	46.9	66.1	60.0	57.7
Qwen3-Omni	30B-A3B	A + V	32.0	13.6	45.6	67.5	74.8	72.3	71.5
video-SALMONN-2	7B	A + V	21.2	17.6	38.8	61.8	71.4	68.5	67.2
UGC-Captioner	3B	A + V	31.6	17.0	48.6	61.4	58.4	57.5	59.1
AVoCaDO	7B	A + V	21.1	16.2	37.3	73.0	74.6	71.8	73.2
TCA-Captioner	30B-A3B	A + V	**18.6**	12.9	**31.5**	**73.8**	**76.0**	**72.9**	**74.2**

– **Temporal F_1:** For each triple in the GT temporal list, the judge assigns a three-way verdict: *correct* (both events present with a consistent temporal link), *incorrect* (both events present but the temporal relationship is clearly reversed or contradicted), or *skipped* (one or both events absent). This three-way protocol decouples temporal *accuracy* from event *coverage*: omitting an event yields a *skipped* verdict rather than a false temporal error. We compute:

$$\text{Prec} = \frac{|\text{correct}|}{|\text{correct}| + |\text{incorrect}|}, \quad \text{Cov} = \frac{|\text{correct}| + |\text{incorrect}|}{|\text{GT}|} \tag{1}$$

$$F_1 = \frac{2 \cdot \text{Prec} \cdot \text{Cov}}{\text{Prec} + \text{Cov}} \tag{2}$$

We report F_1 separately for *Sequential* and *Simultaneous* relations.

5 Experiment

5.1 Caption Evaluation on Public Benchmarks

We first evaluate the audiovisual captioning proficiency of TCA-Captioner across two public benchmarks: the Video-SALMONN-2 [49] and the UGC-VideoCap [58] benchmark. To ensure a fair and consistent comparison, we strictly adhere to the evaluation protocol established by AVoCaDO [7], employing GPT-4.1 as the judge model for Video-SALMONN-2 and GPT-4o for UGC-VideoCap. Our model is benchmarked against a comprehensive suite of representative baselines, including prominent open-source models, such as HumanOmniV2 [61], ARC-Hunyuan-Video [21], MiniCPM-o-2.6 [62], Qwen2.5-Omni [59], Qwen3-Omni [60], video-SALMONN-2 [49], UGC-Captioner [58], and the latest AVoCaDO [7], as well as the powerful closed-source Gemini series [22]. As illustrated

Table 4: Caption evaluation on our TCA-Bench. We employ GPT-4.1 as the judge model. Note that all metric scores have been normalized to a scale of 0–100.

Model	Audio			Visual				AV Binding			AV Temporal		
	Trans. ↑	Ton. ↑	Avg. ↑	Sub. ↑	Sce. ↑	Cine. ↑	Avg. ↑	Char. ↑	Envi. ↑	Total ↑	Seq. ↑	Sim. ↑	Total ↑
Gemini-2.5-Pro	64.2	66.3	65.3	66.0	69.5	63.1	66.2	82.6	47.6	74.3	78.5	76.9	78.0
Gemini-2.5-Flash	62.2	64.6	63.4	66.0	69.6	67.2	67.6	77.3	41.1	68.6	73.3	71.6	72.7
Gemini-3.0-Pro	71.5	68.9	70.2	61.5	62.8	58.5	60.9	87.3	42.8	76.8	80.9	78.3	80.0
Gemini-3.0-Flash	61.7	59.6	60.7	60.0	63.2	50.1	57.8	81.3	34.3	70.1	72.4	71.4	72.0
Qwen2.5-Omni-3B	29.3	32.3	30.8	46.9	52.1	29.1	42.7	41.6	17.2	35.8	33.3	38.0	35.0
Qwen2.5-Omni-7B	33.6	37.8	35.7	51.1	54.6	29.9	45.2	39.9	18.3	34.8	37.5	43.3	39.7
Qwen3-Omni	35.2	42.8	39.0	53.3	61.7	39.2	51.4	48.8	22.2	42.5	46.7	49.9	47.9
UGC-Captioner	39.8	48.2	44.0	48.7	58.8	35.0	47.5	51.5	22.3	44.6	46.1	51.0	47.9
AVoCaDO	59.2	**62.3**	60.8	59.3	63.9	**65.2**	62.8	76.7	36.8	67.2	68.8	67.3	68.3
TCA-Captioner	**61.6**	60.8	**61.2**	**59.6**	**68.9**	62.2	**63.6**	**82.1**	**44.6**	**73.2**	**76.7**	**77.1**	**76.9**

in Table 3, TCA-Captioner achieves state-of-the-art (SOTA) performance among all open-source models and demonstrates formidable competitiveness even when positioned against top-tier closed-source systems.

5.2 Caption Evaluation on Our TCA-Bench

We conduct extensive experiments on our proposed TCA-Bench to evaluate the multi-modal proficiency of TCA-Captioner against a broad spectrum of models, including the closed-source Gemini series (v2.5 and v3.0) and various leading open-source MLLMs. The performance is measured across four primary dimensions: Audio (Transcription, Tone & Emotion), Visual (Subject & Action, Scene & Atmosphere, Cinematography), Audiovisual (AV) Binding (Character, Environment), and AV Temporal (Sequential, Simultaneous). The experimental results in Table 4 reveal several key insights:

Gap between Open and Closed-Source Models: A significant performance disparity exists between existing open-source models and the closed-source Gemini series. This gap is particularly pronounced in the AV Binding and AV Temporal tracks, where open-source models struggle to accurately associate acoustic events with their corresponding visual sources or maintain the logical chronological order of cross-modal interactions.

Evolution of Gemini Series: Gemini-3.0-Pro demonstrates substantial improvements in Audio (70.2 avg.), AV Binding (76.8 total) and AV Temporal (80.0 total) compared to its predecessor, Gemini-2.5-Pro (65.3, 74.3 and 78.0, respectively). However, a performance degradation is observed in the Visual dimension, where its average score drops from 66.2 to 60.9, suggesting a trade-off in its visual perception tuning.

Superiority of TCA-Captioner: Our model consistently outperforms all existing open-source models across both the average and total metrics within the Audio, Visual, AV Binding, and AV Temporal dimensions. Notably, in the AV Binding (73.2) and AV Temporal (76.9) categories, TCA-Captioner exhibits a profound understanding of cross-modal interactions, nearly doubling the performance of the baseline Qwen3-Omni.

Table 5: Effect of different base models and DPO.

Model	Audio	Visual	AV Binding	AV Temporal
Qwen2.5-Omni (SFT)	59.3	61.7	72.5	73.7
Qwen2.5-Omni (SFT+DPO)	62.0	63.1	73.6	75.9
Qwen3-Omni-30B (SFT)	61.2	63.6	73.2	76.9
Qwen3-Omni-30B (SFT+DPO)	63.0	66.8	75.3	79.2

Competitive Edge against Closed-Source Models: Compared to the Gemini family, TCA-Captioner surpasses Gemini-3.0-Flash in every evaluated dimension. Furthermore, our model achieves a superior Video average score (63.6) compared to Gemini-3.0-Pro (60.9). This advantage is primarily attributed to our specialized Visual Checker module, which leverages the high-fidelity perception of Doubao-1.8 to rectify hallucinations and supplement granular visual details that generic models often overlook.

5.3 Exploring Another Base Models and DPO

As shown in Table 5, we conduct additional experiments based on Qwen2.5-Omni, which similarly achieves competitive performance. Furthermore, we construct 2,000 DPO preference pairs based on our OCC pipeline. As demonstrated in Table 5, applying DPO to our SFT captioner significantly enhances performance across all dimensions. This validates the high quality of the OCC refinement data and proves its strong potential for automated preference alignment.

5.4 Analysis for Our OCC Framework

Since the generation of LLM drafts in our benchmark construction utilized the OCC framework, a direct comparison within the full TCA-Bench would introduce inherent bias. To conduct a more rigorous and equitable ablation study of the OCC framework's contributions, we constructed an additional mini-bench comprising 20 samples. These samples utilized Gemini-2.5-Pro for initial draft generation and underwent stringent human auditing and manual refinement to ensure the same structural integrity as the main TCA-Bench. As illustrated in Figure 5, we compared the OCC framework against Gemini-3.0-Pro. The results demonstrate significant improvements across

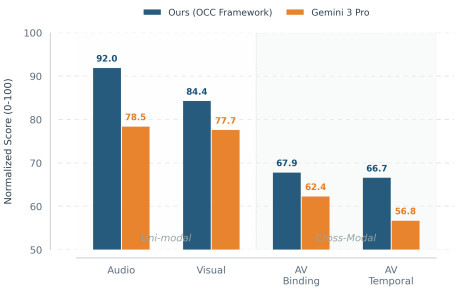


Fig. 5: The performance analysis for OCC framework.

Table 6: Ablation study for the OCC framework.

Model	Audio	Visual	AV Binding	AV Temporal
Observer Only	57.6	56.1	66.5	69.8
Observer + Audio Checker	61.5	57.0	68.2	68.1
Observer + Visual Checker	58.8	64.8	67.8	69.5
Full OCC Framework (A/V-Clip-Checker)	61.2	63.6	73.2	76.9

all evaluated dimensions. The visual total score increased from 77.7 to 84.4, underscoring the critical role of the Visual Checker in refining visual details. The audio total score rose from 78.5 to 92.0, highlighting the substantial impact of the Audio Checker in rectifying auditory content. While AV Temporal alignment saw a robust 9.9 improvement, the gain in AV Binding was more modest at 5.5. This discrepancy suggests that for complex multi-person dialogue cases, severe hallucinations, particularly regarding the attribution of ambient sounds, persist even when the video is segmented into finer clips. Moreover, Table 6 presents an ablation study where models are finetuned on data generated by specific OCC components. The results confirm that the Audio and Visual Checkers significantly boost their respective scores, while the A/V-Clip-Checker markedly enhances AV Binding and Temporal capabilities.

6 Ethical Statement

Our dataset and benchmark comprise publicly available YouTube videos (e.g., variety shows and films) paired with precise timestamps and fine-grained annotations. The suite will be released under a CC BY-NC-SA 4.0 license, strictly restricted to non-commercial research and explicitly prohibiting biometric or surveillance applications. Furthermore, we remain committed to privacy rights; any featured individual may request content removal, which we will fulfill by promptly excluding the relevant segments from future releases.

7 Conclusion

In this work, we present TCA-Captioner, a specialized framework designed to bridge the gap in audiovisual video captioning. By integrating the iterative OCC framework and training on a curated high-density interaction dataset, our approach effectively alleviates modality detachment and temporal incoherence. Furthermore, the introduction of TCA-Bench enables a granular diagnostic evaluation of cross-modal interaction. Experimental results confirm that TCA-Captioner significantly advances the state-of-the-art in generating precise, temporally-aligned, and contextually rich audiovisual narratives.

Acknowledgments. This work was supported by Gusu Innovation and Entrepreneur Leading Talents: No. ZXL2024362, National Natural Science Foundation of China : No. 62406135, Natural Science Foundation of Jiangsu Province: BK20241198, and Meituan.

References

1. Abdar, M., Kollati, M., Kuraparthi, S., Pourpanah, F., McDuff, D., Ghavamzadeh, M., Yan, S., Mohamed, A., Khosravi, A., Cambria, E., et al.: A review of deep learning for video captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Chen, H., Zhang, X., Sun, X., Mingqing, X.: Calibrated harmonic overlaid implicit neural representations for multi-dimensional data. arXiv preprint arXiv:2606.26763 (2026)
4. Chen, L., Wang, J., Pan, Z., Zhu, B., Yang, X., Zhang, C.: Detail++: Training-free detail enhancer for t2i diffusion models. IEEE Transactions on Image Processing **35**, 5982–5994 (2026)
5. Chen, L., You, T., Liu, H., Bao, Z., Jiao, J., Han, X., Ou, Z., Sun, T., Mou, X., Jin, X., et al.: Echo: Efficient chest x-ray report generation with one-step block diffusion. arXiv preprint arXiv:2604.09450 (2026)
6. Chen, L., Wei, X., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Tang, Z., Yuan, L., et al.: Sharegpt4video: Improving video understanding and generation with better captions. Advances in Neural Information Processing Systems **37**, 19472–19495 (2024)
7. Chen, X., Ding, Y., Lin, W., Hua, J., Yao, L., Shi, Y., Li, B., Zhang, Y., Liu, Q., Wan, P., et al.: Avocado: An audiovisual video captioner driven by temporal orchestration. arXiv preprint arXiv:2510.10395 (2025)
8. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
9. Chen, Z., Gao, R., Xiang, T.Z., Lin, F.: Diffusion model for camouflaged object detection. arXiv preprint arXiv:2308.00303 (2023)
10. Chen, Z., Li, Y., Wang, H., Chen, Z., Jiang, Z., Li, J., Wang, Q., Yang, J., Tai, Y.: Ragd: Regional-aware diffusion model for text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19331–19341 (2025)
11. Chen, Z., Zhu, J., Chen, X., Zhang, J., Chen, J., Zeng, Z., Zhang, W., Wang, C., Yang, J., Tai, Y.: L2p: Unlocking latent potential for pixel generation. arXiv preprint arXiv:2605.12013 (2026)
12. Chen, Z., Zhu, J., Chen, X., Zhang, J., Hu, X., Zhao, H., Wang, C., Yang, J., Tai, Y.: Dip: Taming diffusion models in pixel space. arXiv preprint arXiv:2511.18822 (2025)
13. Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., et al.: Qwen2-audio technical report. arXiv preprint arXiv:2407.10759 (2024)
14. Company, T.D.: Tiktok-10m: A large-scale short video dataset for video understanding (2025), https://huggingface.co/datasets/The-data-company/TikTok-10M, accessed: June 30, 2026
15. Du, S., Zou, Y., Li, J., Liu, M., Li, Y., Shang, C., Shen, Q.: Pansharpening for thin-cloud contaminated remote sensing images: a unified framework and benchmark dataset. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 3696–3704 (2026)

16. Du, S., Zou, Y., Wang, Z., Li, X., Li, Y., Shang, C., Shen, Q.: Unsupervised hyperspectral image super-resolution via self-supervised modality decoupling. International Journal of Computer Vision **134**, 152 (2026)
17. Du, Y., Lin, Z., Song, K., Wang, B., Zheng, Z., Ge, T., Zheng, B., Jin, Q.: Vc4vg: Optimizing video captions for text-to-video generation. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 1124–1138 (2025)
18. Fan, T., Nan, K., Xie, R., Zhou, P., Yang, Z., Fu, C., Li, X., Yang, J., Tai, Y.: Instancecap: Improving text-to-video generation via instance-aware structured caption. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28974–28983 (2025)
19. Farré, M., Marafioti, A., Tunstall, L., Von Werra, L., Wolf, T.: Finevideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo (2024), accessed: June 30, 2026
20. Fu, C., Lin, H., Long, Z., Shen, Y., Dai, Y., Zhao, M., Zhang, Y.F., Dong, S., Li, Y., Wang, X., et al.: Vita: Towards open-source interactive omni multimodal llm. arXiv preprint arXiv:2408.05211 (2024)
21. Ge, Y., Ge, Y., Li, C., Wang, T., Pu, J., Li, Y., Qiu, L., Ma, J., Duan, L., Zuo, X., et al.: Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts. arXiv preprint arXiv:2507.20939 (2025)
22. Gemini-3-Pro Team: Gemini-3-Pro. https://gemini.google.com, accessed: June 30, 2026
23. Guo, Y., Gan, Q., Zhang, Y., Liu, J., Hu, Y., Xie, P., Qian, D., Zhang, Y., Li, R., Zhang, Y., et al.: Alive: Animate your world with lifelike audio-video generation. arXiv preprint arXiv:2602.08682 (2026)
24. Huang, G., Mao, J., Huang, F., Liu, F., Luo, X., Liang, Y., Lu, J., Wang, X., Liu, P., Fu, R., Huang, R., Huang, S.L.: Exposure bias can alleviate itself via directional and frequency rectification in flow matching (2026)
25. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
26. Jiang, T., Wu, L., Xia, S., Li, S., Yan, Z., Yang, H., Qiao, Y., Wang, Y.: Imagine before you predict: Interleaved latent visual reasoning for video event prediction. arXiv preprint arXiv:2606.05769 (2026)
27. Jiang, T., Xia, S., Xu, Y., Wu, L., Zeng, X., Wang, L., Qiao, Y., Wang, Y.: Vknowu: Evaluating visual knowledge understanding in multimodal llms. arXiv preprint arXiv:2511.20272 (2025)
28. Kim, Y., Abdelrahman, A.S., Abdel-Aty, M.: Vru-accident: A vision-language benchmark for video question answering and dense captioning for accident scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 761–771 (2025)
29. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
30. Li, X., Du, S., Zou, Y., Xu, H., Jiang, Z., Liu, J.: Unifusion: A unified image fusion framework with robust representation and source-aware preservation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 33869–33880 (2026)
31. Lian, L., Ding, Y., Ge, Y., Liu, S., Mao, H., Li, B., Pavone, M., Liu, M.Y., Darrell, T., Yala, A., et al.: Describe anything: Detailed localized image and video cap-

tioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21766–21777 (2025)
32. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
33. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pretraining for visual language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26689–26699 (2024)
34. Lin, K., Li, L., Lin, C.C., Ahmed, F., Gan, Z., Liu, Z., Lu, Y., Wang, L.: Swinbert: End-to-end transformers with sparse attention for video captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17949–17958 (2022)
35. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
36. Liu, M., Zhang, D., Liu, J., Cui, J., Xie, H., Chen, G., Ye, H., Yang, M.Y., Nex, F., Cheng, H.: Driveva: Video action models are zero-shot drivers. arXiv preprint arXiv:2604.04198 (2026)
37. Liu, Y., Li, S., Liu, Y., Wang, Y., Ren, S., Li, L., Chen, S., Sun, X., Hou, L.: Tempcompass: Do video llms really understand videos? In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 8731–8772 (2024)
38. Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., Kembhavi, A.: Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26439–26455 (2024)
39. Lu, S., Lian, Z., Zhou, Z., Zhang, S., Zhao, C., Kong, A.W.K.: Does flux already know how to perform physically plausible image composition? arXiv preprint arXiv:2509.21278 (2025)
40. Lu, S., Liu, Y., Kong, A.W.K.: Tf-icon: Diffusion-based training-free cross-domain image composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2294–2305 (2023)
41. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.K.: Mace: Mass concept erasure in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6430–6440 (2024)
42. Lu, S., Zhou, Z., Lu, J., Zhu, Y., Kong, A.W.K.: Robust watermarking using generative priors against image editing: From benchmarking to advances. arXiv preprint arXiv:2410.18775 (2024)
43. Ma, Z., Xu, R., Xing, Z., Chu, Y., Wang, Y., He, J., Xu, J., Heng, P.A., Yu, K., Lin, J., et al.: Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. arXiv preprint arXiv:2510.12720 (2025)
44. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In: International Conference on Learning Representations. vol. 2025, pp. 1045–1064 (2025)
45. Nguyen, L.T.P., Yu, Z., Hang, S.L.Y., An, S., Lee, J., Ban, Y., Chung, S., Nguyen, T.H., Maeng, J., Lee, S., et al.: See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. arXiv preprint arXiv:2512.02231 (2025)
46. Qasim, I., Horsch, A., Prasad, D.: Dense video captioning: A survey of techniques, datasets and evaluation protocols. ACM Computing Surveys **57**(6), 1–36 (2025)

47. Shang, Y., Gao, C., Li, N., Li, Y.: A large-scale dataset with behavior, attributes, and content of mobile short-video platform. In: Companion Proceedings of the ACM on Web Conference 2025. pp. 793–796 (2025)
48. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: Pandagpt: One model to instruction-follow them all. In: Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants! pp. 11–23 (2023)
49. Tang, C., Li, Y., Yang, Y., Zhuang, J., Sun, G., Li, W., Ma, Z., Zhang, C.: video-salmonn 2: Caption-enhanced audio-visual large language models. arXiv preprint arXiv:2506.15220 (2025)
50. Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., Zhang, C.: Salmonn: Towards generic hearing abilities for large language models. arXiv preprint arXiv:2310.13289 (2023)
51. Team, M.L., Cai, X., Huang, Q., Kang, Z., Li, H., Liang, S., Ma, L., Ren, S., Wei, X., Xie, R., et al.: Longcat-video technical report. arXiv preprint arXiv:2510.22200 (2025)
52. Team, S.: Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier (2024)
53. Varghese, R., Sambath, M.: Yolov8: A novel object detection algorithm with enhanced performance and robustness. In: 2024 International conference on advances in data engineering and intelligent computing systems (ADICS). pp. 1–6. IEEE (2024)
54. Wang, J., Yuan, L., Zhang, Y., Sun, H.: Tarsier: Recipes for training and evaluating large video description models. arXiv preprint arXiv:2407.00634 (2024)
55. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
56. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
57. Wang, X., Hua, J., Lin, W., Zhang, Y., Zhang, F., Wu, J., Zhang, D., Nie, L.: Haic: Improving human action understanding and generation with better captions for multi-modal large language models. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 10158–10181 (2025)
58. Wu, P., Liu, Y., Zhu, Z., Zhou, E., Shen, J.: Ugc-videocaptioner: An omni ugc video detail caption model and new benchmarks. arXiv preprint arXiv:2507.11336 (2025)
59. Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., Zhang, B., Wang, X., Chu, Y., Lin, J.: Qwen2.5-omni technical report (2025)
60. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al.: Qwen3-omni technical report. arXiv preprint arXiv:2509.17765 (2025)
61. Yang, Q., Yao, S., Chen, W., Fu, S., Bai, D., Zhao, J., Sun, B., Yin, B., Wei, X., Zhou, J.: Humanomniv2: From understanding to omni-modal reasoning with context. arXiv preprint arXiv:2506.21277 (2025)
62. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
63. Yuan, L., Wang, J., Sun, H., Zhang, Y., Lin, Y.: Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. arXiv preprint arXiv:2501.07888 (2025)

64. Zhan, J., Dai, J., Ye, J., Zhou, Y., Zhang, D., Liu, Z., Zhang, X., Yuan, R., Zhang, G., Li, L., et al.: Anygpt: Unified multimodal llm with discrete sequence modeling. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 9637–9662 (2024)
65. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
66. Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., Qiu, X.: Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 15757–15773 (2023)
67. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations. pp. 543–553 (2023)
68. Zhang, L., Meng, W., Zhong, Y., Kong, B., Xu, M., Du, J., Wang, X., Wang, R., Liu, L.: U-cope: Taking a further step to universal 9d category-level object pose estimation. In: European Conference on Computer Vision. pp. 254–270. Springer (2025)
69. Zhang, L., Meng, X., Liu, L., Jiang, H., Wang, J., Wang, R., Lu, C., Liu, J., Zhang, H.: Probing effective and efficient category-level articulated object pose perception. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)
70. Zhang, L., Xu, M., Wang, J., Yu, Q., Yang, L., Li, Y., Lu, C., Wang, R., Liu, L.: Gapt-dar: Category-level garments pose tracking via integrated 2d deformation and 3d reconstruction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22638–22647 (2025)
71. Zhang, L., Zhong, Y., Wang, J., Min, Z., Liu, L., et al.: Rethinking 3d convolution in ℓ_p-norm space. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
72. Zhang, S., Guo, S., Fang, Q., Zhou, Y., Feng, Y.: Stream-omni: Simultaneous multimodal interactions with large language-vision-speech model. arXiv preprint arXiv:2506.13642 (2025)
73. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
74. Zhao, C., Cai, W., Dong, C., Hu, C.: Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2024)
75. Zhao, C., Chen, J., Li, H., Kang, Z., Lu, S., Wei, X., Zhang, K., Yang, J., Tai, Y.: Luve: Latent-cascaded ultra-high-resolution video generation with dual frequency experts. arXiv preprint arXiv:2602.11564 (2026)
76. Zhao, C., Ci, E., Xu, Y., Fan, T., Guan, S., Ge, Y., Yang, J., Tai, Y.: Ultrahr-100k: Enhancing uhr image synthesis with a large-scale high-quality dataset. Advances in Neural Information Processing Systems (2025)
77. Zhao, C., Dong, C., Cai, W., Wang, Y.: Learning a physical-aware diffusion model based on transformer for underwater image enhancement. IEEE Transactions on Geoscience and Remote Sensing (2026)
78. Zhao, C., Xu, Y., Chen, Z., Gu, E., Zhang, K., Liu, X., Yang, J., Tai, Y.: From zero to detail: A progressive spectral decoupling paradigm for uhd image restoration with new benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–18 (2026)

79. Zhong, C., Hou, Q., Zhou, Z., Hao, S., Lu, H., Zhang, Y., Tang, H., Bai, X.: Owlcap: Harmonizing motion-detail for video captioning via hmd-270k and caption set equivalence reward. arXiv preprint arXiv:2508.18634 (2025)
80. Zhou, D., Huang, X., Wang, X., Xie, J., Zhang, Y., Li, L., Li, K., Yang, Z., Yang, Y.: Metapoint: Unlocking precise spatial control in agentic visual generation. arXiv preprint arXiv:2606.05031 (2026)
81. Zhou, D., Li, M., Yang, Z., Lu, Y., Xu, Y., Wang, Z., Huang, Z., Yang, Y.: Bidedpo: Conditional image generation with simultaneous text and condition alignment. arXiv preprint arXiv:2511.19268 (2025)
82. Zhou, D., Li, M., Yang, Z., Yang, Y.: Dreamrenderer: Taming multi-instance attribute control in large-scale text-to-image models. arXiv preprint arXiv:2503.12885 (2025)
83. Zhou, D., Xie, J., Yang, Z., Yang, Y.: 3dis: Depth-driven decoupled instance synthesis for text-to-image generation. arXiv preprint arXiv:2410.12669 (2024)
84. Zhou, X., Arnab, A., Buch, S., Yan, S., Myers, A., Xiong, X., Nagrani, A., Schmid, C.: Streaming dense video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18243–18252 (2024)
85. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)