# DiaDem: Advancing Dialogue Descriptions in Audiovisual Video Captioning for Multimodal Large Language Models

Xinlong Chen[1,2,3⋆], Weihong Lin[3], Jingyun Hua[3], Linli Yao[4], Yue Ding[1,2], Bozhou Li[4], Bohan Zeng[4], Yang Shi[4], Qiang Liu[1,2✉], Yuanxing Zhang[3], Pengfei Wan[3], and Liang Wang[1,2]

[1] NLPR, MAIS, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Kling Team, Kuaishou Technology
[4] Peking University
chenxinlong2025@ia.ac.cn, qiang.liu@nlpr.ia.ac.cn

**Abstract.** Accurate dialogue description in audiovisual video captioning is crucial for downstream understanding and generation tasks. However, existing models generally struggle to produce captions that faithfully reflect spoken content and speaker dynamics. To mitigate this limitation, we propose **DiaDem**, a powerful audiovisual video captioning model capable of generating captions with more precise dialogue descriptions while maintaining strong overall performance. We first design a dedicated pipeline to synthesize high-quality dialogue-aware audiovisual captions for supervised fine-tuning (SFT), and then introduce a difficulty-partitioned two-stage reinforcement learning (RL) strategy to further enhance dialogue descriptions. To enable systematic evaluation of dialogue description capabilities, we present **DiaDemBench**, a comprehensive benchmark designed to evaluate models across diverse dialogue scenarios, focusing on both speaker attribution accuracy and utterance transcription fidelity in audiovisual captions. Extensive experiments on DiaDemBench reveal even commercial models still exhibit substantial room for improvement in dialogue-aware captioning. Notably, DiaDem not only outperforms the Gemini series in dialogue description accuracy but also achieves competitive performance on general audiovisual captioning benchmarks, demonstrating its overall effectiveness. Our project is available at https://diadem-captioner.github.io/.

**Keywords:** Audiovisual · Dialogue · Caption

## 1 Introduction

Audiovisual joint video captioning has recently garnered increasing attention [43, 50]. Compared to visual-only video captioning [6, 45, 59, 64], incorporating auditory signals enables models to generate more comprehensive and human-aligned

⋆ Work done during the author's internship at Kling Team, Kuaishou Technology.

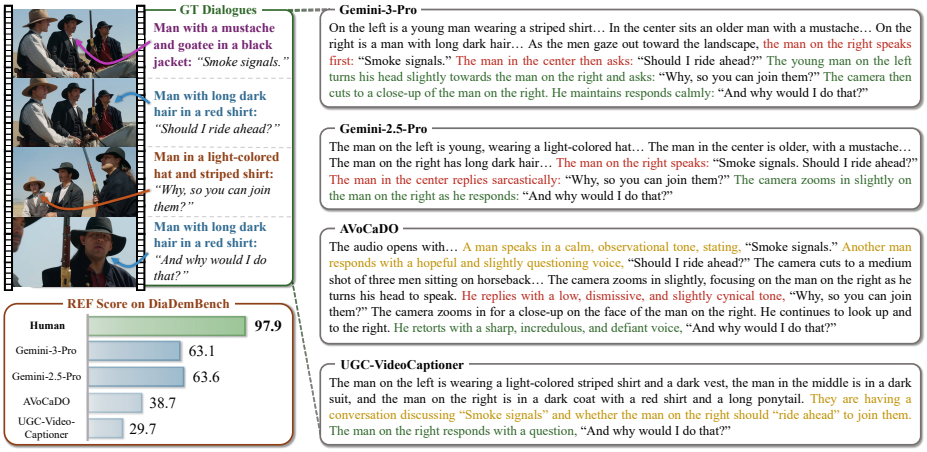


**Fig. 1:** Qualitative and quantitative analysis of dialogue description capabilities in existing models. In the generated audiovisual captions, speaker attributions are color-coded as incorrect, ambiguous, and correct. The REF score quantifies speaker–utterance attribution accuracy, while the captioning examples offer qualitative evidence, jointly indicating substantial room for improvement in current models.

textual descriptions [7, 25]. High-quality audiovisual captions not only facilitate effective alignment across audio, visual, and textual modalities during pretraining [48, 61], but also provide critical support for downstream audiovisual understanding and generation tasks [15, 36, 60].

Despite notable progress in audiovisual video captioning, most existing models and benchmarks primarily prioritize descriptive completeness [43, 50], while overlooking a critical dimension in audiovisual scenarios: *the accuracy of dialogue description, encompassing precise utterance transcription and correct speaker attribution*. As illustrated in Fig. 1, current models consistently struggle to discern "who said what" in challenging multi-party dialogue scenarios. However, faithful speaker-utterance identification and attribution in audiovisual captioning are indispensable for accurately capturing characters' intentions, inferring interpersonal dynamics, and reconstructing the narrative's logical structure.

To enable systematic and fine-grained evaluation of dialogue description capabilities in audiovisual captioning, we first introduce **DiaDemBench**, which comprises 1,039 manually annotated videos featuring diverse dialogue settings, covering single- and multi-shot scenes, varying speaker counts, and challenging scenarios with overlapping speech, thereby providing a comprehensive testbed for evaluating both utterance transcription fidelity and speaker attribution accuracy in audiovisual captioning. As shown in Fig. 1, even advanced open-source and commercial models still exhibit substantial room for improvement in speaker-utterance matching, with a notable performance gap persisting between open-source and commercial models.

To address this issue, we further propose **DiaDem**, an audiovisual video captioning model capable of producing more precise utterance transcription and speaker attribution, while preserving the completeness of other audiovisual details. Building upon AVoCaDO [7], which possesses strong capability in holistic audiovisual captioning, DiaDem is enhanced through targeted post-training to improve dialogue understanding. Specifically, leveraging the complementary strengths of different models, we design a dedicated pipeline to construct a training corpus comprising 70K high-quality audiovisual captions with precise dialogue descriptions, together with 15K general-purpose captions from non-dialogue scenes for SFT, equipping the model with foundational dialogue description skills while maintaining its general captioning performance. Furthermore, we manually annotate 3K dialogue-centric samples and incorporate them into a difficulty-partitioned two-stage RL (specifically, GRPO) strategy, further strengthening the model's ability to transcribe utterances and attribute them to the correct speakers in audiovisual captions. Experimental results show that DiaDem not only outperforms the Gemini series in dialogue description accuracy, but also achieves competitive performance in general audiovisual captioning.

Our contributions are summarized as follows:

- We propose DiaDem, a powerful audiovisual video captioning model that delivers superior dialogue description capability compared to the Gemini series, while achieving competitive general audiovisual captioning quality.
- We introduce DiaDemBench, the first benchmark designed to evaluate the accuracy of dialogue descriptions in audiovisual models, covering both utterance transcription and speaker attribution.
- We design a post-training strategy that first integrates the complementary strengths of different models to synthesize dialogue-aware captions for SFT, followed by a difficulty-partitioned two-stage RL strategy to sharpen more precise dialogue description and boost the overall captioning performance.

## 2 Related Works

### 2.1 Audiovisual Video Captioning with MLLMs

With the rapid advancement of audiovisual understanding models [10, 19, 20, 32, 40, 42, 44, 57], audiovisual captioning models and benchmarks have also evolved accordingly. For instance, video-SALMONN-2 [43] introduces MrDPO to mitigate information loss and hallucinations, and proposes a testset that evaluates captions based on the accuracy and completeness of atomic events. UGC-VideoCaptioner [50] enhances captioning performance by distilling knowledge from Gemini-2.5-Flash [12] combined with reinforcement learning, and further introduces UGC-VideoCap, a benchmark where a judge model scores captions across multiple dimensions. AVoCaDO [7] adopts a two-stage post-training strategy that explicitly emphasizes the temporal alignment of audiovisual events to improve caption fidelity. Qwen3-Omni-Captioner [25] designs an agentic data generation and cleaning pipeline to construct high-quality caption corpora, and

proposes Omni-Cloze, which evaluates caption quality via a cloze-style multiple-choice proxy task.

However, existing works primarily focus on the completeness of descriptive content, often neglecting the accuracy of dialogue descriptions in audiovisual scenarios, which is critical for downstream understanding and generation tasks. To bridge this gap, we propose DiaDem, a captioning model that generates audiovisual captions with dialogue accuracy surpassing that of the Gemini series. Additionally, we introduce DiaDemBench, a benchmark specifically designed to evaluate the accuracy of dialogue descriptions in audiovisual captions.

### 2.2 Speaker Diarization and Recognition

In the audio-only domain, traditional speaker diarization frameworks [8,26] typically consist of multiple components, including Voice Activity Detection (VAD) to identify speech segments, speaker embedding extraction to capture distinctive speaker characteristics, and clustering to group embeddings for speaker identification. Subsequent studies have further incorporated Automatic Speech Recognition (ASR) modules to obtain speaker identities along with their utterances [13,14,35,47,63]. However, due to the accumulation of errors across modular pipelines, later works have explored post-processing with LLMs [17,33,46] or adopted end-to-end approaches to mitigate error propagation [21,22,24,58,62].

In audiovisual settings, although visual information is incorporated, early studies primarily focus on providing speaker IDs paired with speaking timestamps [5,9,11,23,37,51,52], or leveraging auxiliary visual cues (*e.g.* lip movements) to enhance speech recognition under noisy conditions [1,16,27,31,34]. However, these approaches often overlook the alignment between speaker descriptions and utterance content, a crucial aspect of joint audiovisual understanding. In contrast, DiaDem integrates both audio and visual modalities to generate high-quality audiovisual captions with accurate speaker descriptions and precise utterance recognition.

## 3 DiaDemBench

### 3.1 Overview

DiaDemBench is designed to comprehensively evaluate the accuracy of dialogue descriptions in audiovisual video captions, focusing on two critical yet underrepresented dimensions in existing benchmarks: *correct speaker attribution* and *precise utterance transcription*. To enable quantitative assessment along these axes, we introduce a suite of tailored evaluation metrics, namely the REF score for speaker attribution and the ASR score for transcription fidelity. Moreover, we collect a representative set of dialogue-centric videos spanning diverse scenarios, accompanied by manually curated annotations to ensure both coverage and reliability in evaluation.

### 3.2 Evaluation Protocol

The accuracy of dialogue descriptions comprise two key aspects: *correct speaker attribution* and *precise utterance transcription*. Since speaker attribution is meaningful only when the corresponding utterance is accurately transcribed, we adopt a two-stage evaluation protocol. Specifically, we first align the predicted dialogue tuples and ground-truth tuples, where each tuple comprises a speaker description and the associated spoken content. Alignment is performed based on utterance similarity matching, yielding the utterance transcription accuracy score, denoted as ASR. Subsequently, within the successfully matched tuple pairs, we evaluate speaker consistency to obtain the speaker reference accuracy score, denoted as REF. The detailed procedure is described below.

Given the structural variability of dialogue descriptions in audiovisual captions, we first employ Gemini-2.5-Pro as a parser to extract and structure the dialogue information from each caption into a sequence of tuples. Let $P = [p_1, p_2, \ldots, p_M]$ denote the predicted dialogue sequence and $G = [g_1, g_2, \ldots, g_N]$ denote the ground-truth dialogue sequence, where $M$ and $N$ are the respective sequence lengths. For any dialogue tuple $x_i \in P \cup G$, we denote its utterance as $u(x_i)$ and its speaker as $s(x_i)$.

To quantify the similarity between any pair of utterances $u(p_i)$ and $u(g_j)$, where $i \in [1, M]$ and $j \in [1, N]$, we use the normalized edit distance:

$$\mathrm{Sim}\big(u(p_i), u(g_j)\big) = 1 - \frac{\mathrm{ed}\big(u(p_i), u(g_j)\big)}{\max\big(\big|(u(p_i)\big|, \big|u(g_j)\big|\big)}, \tag{1}$$

where $\mathrm{ed}(\cdot)$ denotes the standard Levenshtein edit distance[5] between two strings.

To achieve optimal matching between the two dialogue sequences, we aim to identify subsequences of equal length from $P$ and $G$ such that (i) each matched utterance pair exceeds a predefined similarity threshold $\gamma$, and (ii) the total utterance similarity across all matched pairs is maximized. This problem can be naturally formulated and solved via dynamic programming [3].

However, a naive one-to-one matching strategy, as used in AVoCaDO, fails to account for segmentation inconsistencies between the predicted and ground-truth dialogues, which may lead to counterintuitive matching results. As illustrated in Fig. 2, in the left example, naive matching causes $p_2$ to be omitted, whereas merging adjacent utterances from the same speaker resolves the issue. However, in the right example, naive matching aligns better with human intuition, while forced merging prevents both $p_1$ and $p_2$ from being successfully matched due to insufficient similarity. Moreover, the segmentation of consecutive utterances from the same speaker is inherently subjective and model-dependent, making it hard to define a canonical segmentation rule.

To alleviate the impact of segmentation variability, we propose an *adaptive merging strategy*. Specifically, during dynamic programming, adjacent utterances from the same speaker are dynamically merged whenever doing so improves alignment fidelity, yielding matches that better reflect human intuition.

[5] https://en.wikipedia.org/wiki/Levenshtein_distance

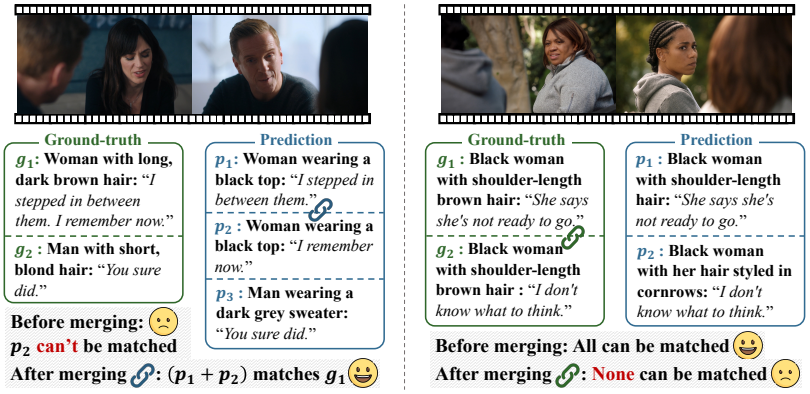


**Fig. 2:** Illustration of potential segmentation inconsistencies between ground-truth annotations and model predictions. Regardless of whether adjacent utterances from the same speaker are merged, suboptimal matching may still occur.

Formally, let $F_{i,j}$ denote the maximum cumulative utterance similarity achievable between the first $i$ elements of $P$ and the first $j$ elements of $G$. The recurrence relation, which incorporates adaptive merging, is defined as follows:

$$F_{i,j} = \begin{cases} 0, & \text{if } i = 0 \text{ or } j = 0, \\ \max\left\{F_{i-1,j}, F_{i,j-1}\right\}, & \text{if } i > 0,\ j > 0, \text{ and } \Phi\big(P^i_{i-k+1},\, G^j_{j-l+1}\big) < \gamma, \\ & \quad \text{for all } 1 \le k, l \le W, \\ \max\Big\{F_{i-1,j}, F_{i,j-1}, & \\ \qquad \max\limits_{1\le k,\, l\le W} \left\{F_{i-k,j-l} + \Phi(P^i_{i-k+1}, G^j_{j-l+1})\right\}\Big\}, & \text{otherwise.} \end{cases} \tag{2}$$

Here, $P^b_a$ and $G^b_a$ denote contiguous subsequences from index $a$ to $b$ in the predicted and ground-truth dialogue sequences, respectively. The hyperparameter $W$ controls the maximum merging window size, balancing matching accuracy against computational efficiency. The function $\Phi(\tilde{P}, \tilde{G})$ computes the utterance similarity between two merged subsequences and is defined as:

$$\Phi(\tilde{P}, \tilde{G}) = \begin{cases} -\infty, & \text{if } \exists\, p_i, p_j \in \tilde{P} \text{ s.t. } s(p_i) \neq s(p_j) \\ & \quad \text{or } \exists\, g_i, g_j \in \tilde{G} \text{ s.t. } s(g_i) \neq s(g_j), \\ \mathrm{Sim}\Big(\mathrm{concat}_{p_i \in \tilde{P}} u(p_i), \mathrm{concat}_{g_j \in \tilde{G}} u(g_j)\Big), & \text{otherwise.} \end{cases} \tag{3}$$

Importantly, merging is restricted to adjacent utterances from the same speaker to avoid introducing speaker ambiguity. Additional implementation details are provided in Sec. A.5 of the supplementary material.

Upon completing utterance matching, we obtain the utterance accuracy score $S_{\text{utterance}} = F_{M,N}$ and the corresponding adaptively merged matched subsequences $P'$ and $G'$. We then compute the speaker reference accuracy score

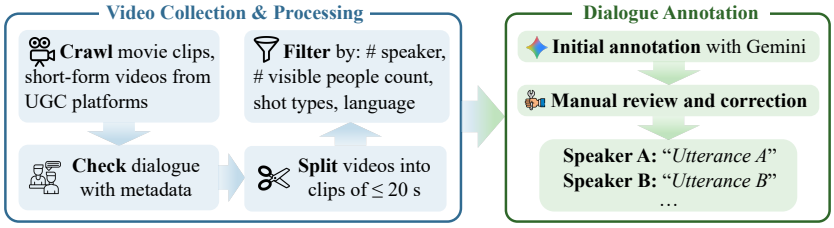


**Fig. 3:** Overview of the data curation pipeline for DiaDemBench.

$S_{\text{speaker}}$ based on speaker consistency within each matched tuple pair. Specifically, we use Gemini-2.5-Flash as the judge model, which, conditioned on the video content $v$, assesses whether the speaker descriptions in each matched tuple pair are consistent and assigns a binary score:

$$S_{\text{speaker}} = \sum_{(p_i, g_i) \in (P', G')} \text{Judge}\big(s(p_i), s(g_i), v\big). \tag{4}$$

To ensure fair comparison across samples with varying dialogue lengths, both $S_{\text{speaker}}$ and $S_{\text{utterance}}$ are normalized. Let $M'$ and $N'$ denote the lengths of the adaptively merged sequences $P'$ and $G'$, respectively. The scores are bounded in $[0, \min(M', N')]$, enabling the computation of precision (normalized by $M'$), recall (normalized by $N'$), and the corresponding F1 score. These F1 scores serve as our final evaluation metrics:

$$\text{REF} = \text{F1}(S_{\text{speaker}}, M', N'), \quad \text{ASR} = \text{F1}(S_{\text{utterance}}, M', N'). \tag{5}$$

Detailed comparisons with prior audio-only multi-speaker transcription metrics [28–30, 41, 49] are provided in Sec. A.6 of the supplementary material.

### 3.3 Data Curation

As shown in Fig. 3, our data curation pipeline begins with the collection of movie clips and short-form videos from public user-generated content (UGC) platforms, targeting materials with rich dialogue scenarios through metadata filtering. Considering the context window limitations of many current models, raw videos are segmented into clips no longer than 20 seconds using PySceneDetect[6], thereby ensuring speaker-related details can be captured at relatively high resolution and frame rate. We then employ multiple open-source models to infer and cross-validate key attributes of each clip, including speaker counts, visible people counts, shot editing types and language information. Based on these attributes, we filter the clips and obtain 1,039 video segments with broad category coverage and balanced distributions.

[6] https://www.scenedetect.com/

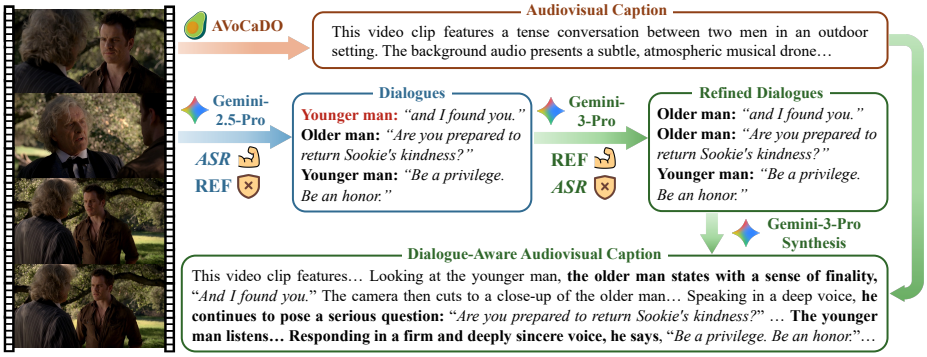


**Fig. 4:** Overview of the data annotation pipeline for SFT. ASR and REF refer to the capabilities of utterance transcription and speaker attribution, respectively.

Each selected clip subsequently undergoes fine-grained dialogue annotation. To improve efficiency, we first generate initial dialogue descriptions using Gemini-2.5-Pro, which are then refined by human annotators to ensure both precise utterance transcriptions and correct speaker attribution. To guarantee high annotation quality, a sample is accepted only when all three annotators reach consensus, as verified by a senior supervisor. Further details are provided in Sec. A.3 of the supplementary material.

## 4 DiaDem

### 4.1 Overview

After constructing DiaDemBench, we evaluate a wide range of models. The results in Tab. 1 indicate a notable performance gap persisting between open-source and commercial models. To mitigate this, we propose DiaDem, a powerful audiovisual video captioning model that integrates a carefully designed data annotation pipeline for SFT and a difficulty-partitioned two-stage RL strategy, thereby achieving superior dialogue description capability.

### 4.2 Data Annotation Pipeline for SFT

Through our empirical case analysis, we observe complementary strengths between two Gemini variants: Gemini-2.5-Pro excels at transcribing utterances, whereas, when given comparable transcription performance, Gemini-3-Pro exhibits superior speaker attribution ability. Capitalizing on this observation, we design a data annotation pipeline, as illustrated in Fig. 4.

First, Gemini-2.5-Pro is employed to generate raw dialogue descriptions with high utterance fidelity but relatively weak speaker attribution. These outputs are then refined by Gemini-3-Pro, which corrects speaker attribution, yielding 70K high-quality dialogue descriptions. In parallel, we utilize the holistic captioning

capability of AVoCaDO to produce base audiovisual captions for these samples. Gemini-3-Pro then integrates the refined dialogue descriptions into these base captions, resulting in 70K high-quality audiovisual captions with more precise dialogue descriptions. In addition, we annotate 15K non-dialogue videos to preserve the model's general audiovisual captioning capability. Details regarding the video sources are provided in Sec. B.1 of the supplementary material.

We then perform SFT on AVoCaDO using the combined 85K dataset, equipping the model with foundational dialogue description skills while maintaining its general captioning performance.

### 4.3 Post-training with RL

**Preliminary** Group Relative Policy Optimization (GRPO) [39] has emerged as a widely adopted RL algorithm, primarily due to its computational efficiency achieved by eliminating the need for a critic model in Proximal Policy Optimization (PPO) [38]. Instead of estimating absolute value functions, GRPO leverages relative rewards within a group of sampled responses to compute advantages. Specifically, for each input query $q$, a group of $G$ responses $\{o_1, o_2, ..., o_G\}$ is sampled from the old policy $\pi_{\theta_{old}}$. The corresponding rewards $\{r_1, r_2, ..., r_G\}$ are then computed to derive the advantage $A_i$ for each response $o_i$:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}. \tag{6}$$

The current policy $\pi_\theta$ is then updated by maximizing the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[ \frac{1}{G} \sum_{i=1}^{G} \left( \min\left( r_i(\theta) A_i, \text{clip}\Big( r_i(\theta), 1-\varepsilon, 1+\varepsilon \Big) A_i \right) - \beta \cdot \mathbb{D}_{\text{KL}}\left( \pi_\theta || \pi_{\text{ref}} \right) \right) \right], \tag{7}$$

where $r_i(\theta) = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}$ denotes the importance sampling ratio, $\epsilon$ is the clipping parameter that restricts policy updates within a trust region, $\beta$ adjusts the strength of the KL divergence penalty, and $\pi_{\text{ref}}$ is the reference policy model employed to enhance training stability.

**Difficulty-Partitioned Two-Stage RL Strategy** Beyond SFT, we further enhance DiaDem through RL, specifically GRPO, with 3K manually annotated high-quality dialogue descriptions. We utilize the sum of REF and ASR scores defined in Eq. 5 as the dialogue reward, combined with the checklist-based reward and length-regularized reward from the base model, to serve as the overall reward function for GRPO, which is detailed in Sec. B.2 of the supplementary material.

As shown in Eqs. 6 and 7, effective policy updates in GRPO rely on reward variations across multiple rollouts of the same sample. However, our pilot experiments reveal that for samples that are either overly simple or excessively difficult, dialogue reward scores exhibit minimal variance across repeated inferences. This

**Table 1:** Model performance on DiaDemBench. $N$ denotes the speaker count. "Overlap" refers to subsets with temporally overlapping speech and is mutually exclusive with the groups defined by speaker count $N$. *For ARC-Qwen-Video-Narrator, speaker and utterance information appears only in the thinking phase rather than the final answer, thus we use the thinking content as the model's output for evaluation.

| Model | Size | Overall (REF \| ASR) | Single-shot (REF \| ASR) | | | | | Multi-shot (REF \| ASR) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | All | $N=1$ | $N=2$ | $N\geq3$ | Overlap | All | $N=1$ | $N=2$ | $N\geq3$ | Overlap |
| Human (avg. of 3 authors) | - | 97.9 \| 97.1 | 98.0 \| 97.4 | 98.7 \| 97.9 | 98.1 \| 97.5 | 97.8 \| 97.3 | 87.3 \| 89.2 | 97.9 \| 96.9 | 98.5 \| 97.4 | 98.4 \| 97.3 | 98.1 \| 97.2 | 88.1 \| 86.7 |
| Gemini-2.5-Pro | - | 63.6 \| 74.8 | 54.2 \| 66.6 | 62.3 \| 71.1 | 59.4 \| 66.7 | **43.8** \| **62.6** | 52.8 \| **80.0** | 69.0 \| 79.6 | 64.5 \| 77.3 | **75.6** \| 81.7 | 67.8 \| 80.3 | 55.4 \| **68.0** |
| Gemini-3-Pro | - | 63.1 \| 71.0 | 51.8 \| 59.8 | 61.2 \| 62.0 | 56.6 \| 61.1 | 40.5 \| 56.5 | 56.3 \| 68.6 | 69.7 \| 77.5 | 70.5 \| 80.7 | 71.6 \| 77.0 | **68.5** \| 76.8 | **57.2** \| 65.1 |
| Gemini-2.5-Flash | - | 58.5 \| 73.3 | 48.2 \| 64.1 | 55.7 \| 65.3 | 51.9 \| 66.2 | 39.4 \| 60.6 | **56.7** \| 77.8 | 64.3 \| 78.6 | 64.5 \| 78.4 | 68.3 \| 78.4 | 62.3 \| 80.3 | 48.4 \| 62.4 |
| video-SALMONN-2 | 7B | 11.5 \| 16.6 | 9.6 \| 13.2 | 18.4 \| 20.1 | 7.2 \| 9.4 | 5.8 \| 12.0 | 7.6 \| 10.3 | 12.6 \| 18.5 | 16.4 \| 18.9 | 11.2 \| 16.1 | 11.1 \| 19.7 | 8.9 \| 17.7 |
| ARC-Qwen-Video | 7B | 11.7 \| 17.2 | 8.2 \| 12.4 | 16.1 \| 15.3 | 7.0 \| 13.2 | 4.0 \| 9.9 | 0.0 \| 5.5 | 13.4 \| 19.6 | 17.5 \| 25.0 | 15.3 \| 20.4 | 9.3 \| 15.8 | 0.0 \| 2.8 |
| HumanOmniV2 | 7B | 15.3 \| 21.5 | 16.2 \| 22.7 | 18.2 \| 24.4 | 17.6 \| 25.7 | 13.4 \| 18.8 | 11.0 \| 12.7 | 14.7 \| 20.6 | 19.9 \| 22.7 | 13.2 \| 20.7 | 12.8 \| 19.8 | 8.6 \| 12.7 |
| OmniVinci | 9B | 17.1 \| 25.2 | 14.4 \| 23.6 | 12.7 \| 19.0 | 21.8 \| 29.8 | 8.1 \| 20.7 | 0.0 \| 0.0 | 18.5 \| 26.0 | 23.9 \| 31.8 | 17.6 \| 25.2 | 16.8 \| 24.2 | 2.9 \| 9.0 |
| Qwen2.5-Omni | 7B | 26.1 \| 37.1 | 26.4 \| 36.8 | 38.6 \| 48.8 | 24.6 \| 33.0 | 18.9 \| 31.5 | 10.2 \| 28.6 | 25.9 \| 37.1 | 35.0 \| 45.7 | 27.7 \| 37.3 | 18.4 \| 32.1 | 11.3 \| 14.6 |
| UGC-VideoCaptioner | 3B | 29.7 \| 47.0 | 31.0 \| 44.3 | 50.8 \| 56.7 | 27.1 \| 40.8 | 20.5 \| 38.7 | 31.0 \| 45.3 | 28.9 \| 48.5 | 41.6 \| 59.5 | 28.6 \| 44.8 | 21.3 \| 45.2 | 12.7 \| 28.2 |
| ARC-Qwen-Video-Narrator* | 7B | 32.8 \| 48.5 | 29.0 \| 41.2 | 43.5 \| 51.9 | 30.0 \| 41.7 | 17.5 \| 33.2 | 4.6 \| 7.8 | 34.7 \| 52.5 | 42.4 \| 62.1 | 37.8 \| 52.9 | 27.6 \| 47.9 | 0.8 \| 2.9 |
| Qwen3-Omni-Instruct | 30B-A3B | 36.8 \| 47.5 | 32.0 \| 40.6 | 43.0 \| 47.8 | 33.3 \| 42.7 | 23.8 \| 34.5 | 0.0 \| 0.0 | 39.2 \| 51.0 | 46.7 \| 57.0 | 39.6 \| 50.5 | 35.1 \| 49.2 | 19.3 \| 25.3 |
| AVoCaDO | 7B | 38.7 \| 51.7 | 33.9 \| 45.4 | 48.7 \| 54.4 | 30.6 \| 41.3 | 27.0 \| 42.6 | 28.6 \| 51.1 | 41.4 \| 55.3 | 39.4 \| 52.1 | 43.5 \| 53.1 | 41.8 \| 60.7 | 31.9 \| 38.3 |
| Qwen3-Omni-Captioner | 30B-A3B | 43.9 \| 58.8 | 41.0 \| 53.5 | 56.1 \| 60.6 | 38.5 \| 51.7 | 32.5 \| 49.7 | 43.6 \| 63.1 | 45.4 \| 61.7 | 49.3 \| 62.2 | 49.6 \| 61.3 | 40.5 \| 62.8 | 29.6 \| 44.7 |
| DiaDem (Ours) | 7B | **65.9** \| **79.3** | **57.2** \| **70.3** | **74.2** \| **79.1** | **60.6** \| **72.4** | 42.4 \| 62.2 | 55.7 \| 68.3 | **70.9** \| **84.4** | **76.9** \| **87.5** | 75.4 \| **86.5** | 65.6 \| **83.2** | 42.9 \| 56.1 |

may lead to uninformative gradients that dilute the learning signal during policy optimization, thereby reducing training efficiency. To mitigate this issue, we first discard samples whose multiple rollouts yield nearly identical rewards due to simplicity, and adopt a difficulty-partitioned two-stage training strategy to progressively strengthen the model's ability to learn from challenging instances.

Specifically, in Stage 1, the model is trained on the entire filtered dataset to improve generalization across samples of varying difficulty. Prior to Stage 2, we partition the filtered dataset based on the average dialogue rewards and isolate a subset of high-difficulty samples. In Stage 2, we double this high-difficulty subset to further strengthen the model's performance on challenging cases. Further implementation details are provided in Sec. B.2 of the supplementary material.

## 5 Experiments

### 5.1 Experimental Settings

We evaluate the dialogue description accuracy in audiovisual captioning of 14 representative models on DiaDemBench, including the Gemini series [12], video-SALMONN-2 [43], OmniVinci [56], the ARC-Qwen-Video series [18], HumanOmniV2 [55], UGC-VideoCaptioner [50], AVoCaDO [7], the Qwen-Omni series [25, 53, 54], as well as our proposed DiaDem.

In addition, we also assess the overall audiovisual captioning performance of DiaDem on the video-SALMONN-2 testset [43] and UGC-VideoCap [50].

### 5.2 Experimental Results

**Evaluation of Dialogue Description Quality** Tab. 1 presents the performance of various models on DiaDemBench in terms of their ability to accurately describe dialogues in audiovisual captions. The results demonstrate that the

**Table 2:** Model performance on the audiovisual video captioning benchmarks. Following AVoCaDO, we replace the judge model for the video-SALMONN-2 testset with GPT-4.1 to ensure more reliable evaluation.

| Model | Size | SALMONN-2 testset | | | UGC-VideoCap | | | |
|---|---|---|---|---|---|---|---|---|
| | | Miss ↓ | Hall. ↓ | Total ↓ | Audio ↑ | Visual ↑ | Detail ↑ | Avg. ↑ |
| Gemini-2.5-Pro | - | 18.1 | 13.3 | 31.3 | 69.5 | 74.7 | 73.7 | 72.6 |
| Gemini-3-Pro | - | 21.4 | 14.2 | 35.6 | 69.7 | 75.6 | 74.3 | 73.2 |
| Gemini-2.5-Flash | - | 19.3 | 13.9 | 33.3 | 69.1 | 75.8 | 74.0 | 73.0 |
| Qwen2.5-Omni | 7B | 41.7 | 15.4 | 57.1 | 46.9 | 66.1 | 60.0 | 57.7 |
| Qwen3-Omni-Instruct | 30B-A3B | 32.0 | **13.6** | 45.6 | 67.5 | 74.8 | 72.3 | 71.5 |
| Qwen3-Omni-Captioner | 30B-A3B | 31.0 | 16.6 | 47.6 | 69.0 | 75.5 | 72.3 | 72.5 |
| UGC-VideoCaptioner | 3B | 31.6 | 17.0 | 48.6 | 61.4 | 58.4 | 57.5 | 59.1 |
| video-SALMONN-2 | 7B | 21.2 | 17.6 | 38.8 | 61.8 | 71.4 | 68.5 | 67.2 |
| AVoCaDO | 7B | **21.1** | 16.2 | 37.3 | 73.0 | 74.6 | 71.8 | 73.2 |
| DiaDem (Ours) | 7B | **21.1** | 15.5 | **36.5** | **75.4** | **76.8** | **74.6** | **75.6** |

closed-source Gemini series substantially outperforms all previous open-source models in both speaker attribution and utterance transcription. However, a considerable gap with human performance remains. Among previous open-source models, the Qwen3-Omni series achieves leading results. AVoCaDO also attains competitive performance due to its holistic audiovisual captioning capabilities.

Notably, our proposed DiaDem, empowered by effective data construction and training strategies, surpasses the best-performing Gemini series by 2.3% in overall speaker attribution and by 4.5% in utterance transcription, and demonstrates consistent superiority in both single-shot and multi-shot scenarios. Nevertheless, under more challenging conditions such as scenes involving three or more speakers or overlapping speech, DiaDem still slightly lags behind Gemini, which constitutes one of our key directions for future improvement.

A detailed analysis and discussion of model behaviors across diverse dimensions of DiaDemBench is provided in Sec. A.2 of the supplementary material.

**Evaluation of Overall Captioning Quality** To verify whether DiaDem improves dialogue description capability without compromising other aspects of audiovisual captioning, we evaluate its overall performance on two benchmarks specifically designed to measure holistic audiovisual captioning quality: the video-SALMONN-2 testset and UGC-VideoCap, as summarized in Tab. 2.

The results show that, while substantially enhancing dialogue description accuracy, DiaDem does not degrade other aspects of audiovisual captioning. On the contrary, it further improves the captioning quality of the base model and even outperforms the strongest Gemini-series model on UGC-VideoCap by 2.4%.

### 5.3 Ablation Studies

**Ablation on the Post-training Pipeline** To evaluate the contribution of each component in our post-training pipeline, we conduct a series of ablation experiments, with results summarized in Tab. 3.

**Table 3:** Ablation study on our post-training pipeline.

| Model | DiaDem-Bench | SALMONN-2 testset ↓ | UGC-VideoCap |
|---|---|---|---|
| AVoCaDO (base model) | 38.7 \| 51.7 | 37.3 | 73.2 |
| *Staged Post-training Recipe* | | | |
| + SFT | 59.3 \| 74.0 | 42.0 | 74.8 |
| + GRPO-Stage-1 | 64.7 \| 77.6 | 38.4 | 74.7 |
| + GRPO-Stage-2 (DiaDem) | **65.9** \| **79.3** | **36.5** | **75.6** |
| *GRPO Variants* | | | |
| + GRPO w/o staged training | 64.5 \| 78.2 | 37.5 | 74.8 |
| + GRPO w/o easy filtering | 63.2 \| 78.2 | 37.7 | 74.5 |
| + GRPO w/o dialogue reward | 60.1 \| 74.3 | 37.0 | 75.2 |
| *w/o GRPO* | | | |
| + SFT using GRPO data | 59.2 \| 73.6 | 42.4 | 75.0 |

Benefiting from our high-quality dialogue-aware audiovisual captioning data construction strategy, the SFT stage substantially enhances the base model's ability to describe dialogues and also improves overall audiovisual caption quality on UGC-VideoCap. However, on the video-SALMONN-2 testset, performance degrades to a level comparable to the SFT version of the base model without GRPO, which scores 41.4 in the original paper. We hypothesize that this degradation arises from a mismatch in training paradigms, where SFT may override the performance gains established during the earlier GRPO stage. Subsequent application of our difficulty-partitioned two-stage GRPO effectively mitigates this issue, yielding steady improvements in both dialogue description accuracy and general audiovisual captioning performance.

To further dissect the design choices within the GRPO stage, we conduct additional ablations: (i) "w/o staged training" merges data from both stages into a single training phase; (ii) "w/o easy filtering" disables the filtering mechanism that discards overly simple samples; and (iii) "w/o dialogue reward" removes our proposed dialogue reward. All three variants underperform the full difficulty-partitioned two-stage GRPO to varying degrees, thereby validating the effectiveness of our training strategy.

Additionally, to investigate whether the performance gains from GRPO are merely attributed to the increased data volume, we adapt the SFT data construction pipeline outlined in Sec. 4.2 to convert the human-annotated dialogue descriptions in the GRPO dataset into dialogue-aware audiovisual captions for additional SFT. However, this modification even leads to a slight performance drop relative to the original SFT model, indicating that the gains from GRPO primarily stem from the reward design and staged training strategy, rather than from exposure to more data.

**Ablation on the Enhanced Dialogue Reward** As discussed in Sec. 3.2, we identify key limitations of the dialogue reward used in AVoCaDO and propose targeted enhancements to address them. To validate the effectiveness of these

**Table 4:** Ablation on our enhanced dialogue reward.

| Model | DiaDem-Bench | SALMONN-2 testset ↓ | UGC-VideoCap |
|---|---|---|---|
| *GRPO w/ the problematic dialogue reward in AVoCaDO* | | | |
| Stage-1 | 63.5 \| 76.9 | 40.1 | 74.6 |
| Stage-2 | 63.7 \| 77.2 | 37.1 | 74.9 |
| *GRPO w/ our enhanced dialogue reward* | | | |
| Stage-1 | 64.7 \| 77.6 | 38.4 | 74.7 |
| Stage-2 (DiaDem) | **65.9** \| **79.3** | **36.5** | **75.6** |

improvements, we conduct an ablation study, with results presented in Tab. 4. The results indicate that our enhanced dialogue reward consistently outperforms the original formulation across both stages of GRPO training, yielding not only more accurate dialogue descriptions but also enhanced overall audiovisual captioning performance.

**Ablation on the Judge Model** In the main experiments, we use Gemini-2.5-Pro to extract structured dialogue descriptions from audiovisual captions, followed by Gemini-2.5-Flash to assess the consistency of speaker descriptions within matched dialogue tuple pairs. To account for scenarios where closed-source model APIs are unavailable, and to further assess the generalizability of our evaluation protocol with respect to the choice of judge model, we replace the Gemini series models with Qwen3-VL-32B [2] and report the results in Tab. 5.

The experimental results reveal that, although the absolute scores produced by different judge models exhibit minor fluctuations, which may stem from inherent model-specific biases, the relative ranking remains largely consistent, with the Pearson correlation [4] achieving 0.995 ($p < 0.001$) for the REF metric and 0.983 ($p < 0.001$) for the ASR metric. This indicates that our evaluation protocol is robust to the choice of judge model. As long as the judge model possesses strong capabilities and can deliver stable and fair judgments, it is suitable for integration into

**Table 5:** Ablation study on the judge model. *In the main experiments, we use two Gemini series models to balance cost and accuracy. In this ablation, both Gemini models are replaced with Qwen3-VL-32B.

| Model | Gemini Series* REF \| ASR | Qwen3-VL-32B REF \| ASR |
|---|---|---|
| Gemini-2.5-Pro | <u>63.6</u> \| <u>74.8</u> | 59.7 \| <u>71.5</u> |
| Gemini-3-Pro | 63.1 \| 71.0 | <u>61.3</u> \| 69.1 |
| Gemini-2.5-Flash | 58.5 \| 73.3 | 58.0 \| <u>71.5</u> |
| video-SALMONN-2 | 11.5 \| 16.6 | 11.0 \| 15.5 |
| Qwen2.5-Omni | 26.1 \| 37.1 | 24.8 \| 34.6 |
| UGC-VideoCaptioner | 29.7 \| 47.0 | 30.5 \| 45.0 |
| ARC-Qwen-Video-Narrator | 32.8 \| 48.5 | 27.6 \| 35.3 |
| Qwen3-Omni-Instruct | 36.8 \| 47.5 | 36.7 \| 46.2 |
| AVoCaDO | 38.7 \| 51.7 | 36.5 \| 50.3 |
| Qwen3-Omni-Captioner | 43.9 \| 58.8 | 43.8 \| 58.7 |
| DiaDem (Ours) | **65.9** \| **79.3** | **63.4** \| **78.3** |

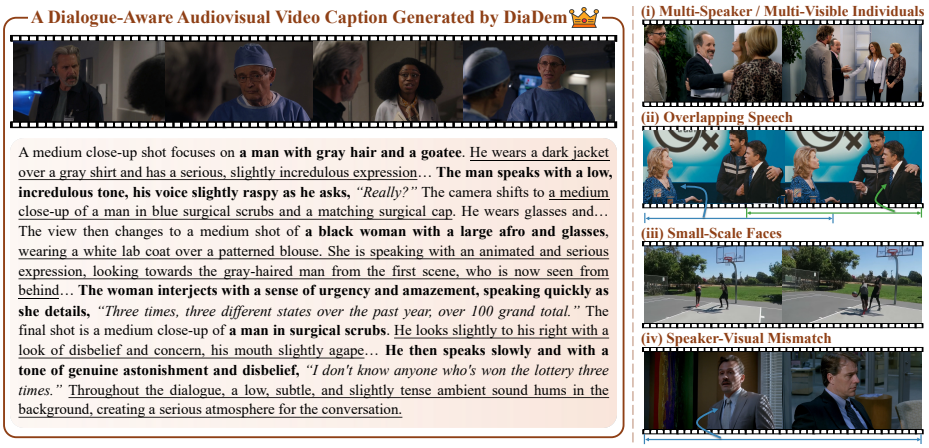


**Fig. 5:** On the left, we present an illustration of an audiovisual video caption with accurate dialogue descriptions generated by DiaDem, featuring both **correct speaker attribution** and *precise utterance transcription*, as well as other general audiovisual details. On the right, we showcase four representative dialogue scenarios from DiaDemBench that are commonly challenging for existing models to produce accurate dialogue descriptions, with the aim of providing insights for future advancements in audiovisual video captioning models.

DiaDemBench. Additionally, aggregating judgments from diverse judge models can help mitigate individual model biases, thereby yielding a more reliable and robust evaluation, albeit at an increased computational cost.

Furthermore, to verify that the ranking aligns with human judgment, we randomly sample 200 instances and manually score model outputs using a five-level scoring scheme: 5 for accuracy $> 95\%$, 4 for accuracy $> 80\%$, 3 for accuracy $> 60\%$, 2 for accuracy $> 40\%$, and 1 for accuracy $\leq 40\%$. As shown in Tab. 6, the capability rankings obtained from human evaluation on both the REF and ASR metrics are highly consistent with those produced by the Gemini-assisted automatic evaluation, yielding Pearson correlation coefficients of 0.994 and 0.982, respectively.

**Table 6:** Correlation with human judgement.

| Model | Gemini | | Human | |
|---|---|---|---|---|
| | REF | ASR | REF | ASR |
| Gemini-2.5-Pro | 63.6 | 74.8 | 3.46 | 3.81 |
| Gemini-3-Pro | 63.1 | 71.0 | 3.38 | 3.64 |
| AVoCaDO | 38.7 | 51.7 | 1.83 | 2.19 |
| Qwen3-Omni-Captioner | 43.9 | 58.8 | 1.96 | 2.33 |
| DiaDem (Ours) | **65.9** | **79.3** | **3.71** | **4.06** |

## 5.4 Qualitative Analysis

In the left panel of Fig. 5, we present an illustrative example of a dialogue-aware audiovisual video caption generated by DiaDem. The result demonstrates Di-

aDem's capability to produce accurate dialogue descriptions with both correct speaker attribution and precise utterance transcription, while simultaneously delivering strong descriptive performance regarding other audiovisual details within the scene. Additional qualitative comparisons between DiaDem and two strong Gemini series models are provided in Sec. C of the supplementary material.

In the right panel of Fig. 5, we highlight four representative dialogue scenarios from DiaDemBench that remain challenging for state-of-the-art audiovisual video captioning models. In addition to the commonly acknowledged difficulties such as scenes with multiple speakers, multiple visible individuals, or overlapping speech, our extensive case studies identify several less-explored yet critical failure cases that frequently lead to erroneous descriptions in current models. These include situations where speakers occupy only small facial regions, as well as cases involving audiovisual mismatches between the off-screen speakers and the characters currently visible on the screen. We hope that these observations can provide valuable insights for the future development of more robust and reliable audiovisual video captioning models.

## 6 Conclusion

This paper focuses on the critical yet underexplored challenge of accurately describing dialogues within audiovisual video captioning. We first introduce a suite of evaluation metrics designed to faithfully quantify two core components of dialogue descriptions: correct speaker attribution and precise utterance transcription. Based on these metrics, we construct DiaDemBench, which features diverse dialogue scenarios and enables a comprehensive and reliable evaluation of dialogue description fidelity in audiovisual captions.

Evaluations on DiaDemBench reveal a substantial performance gap between existing open-source and commercial models. To bridge this gap, we propose DiaDem, an audiovisual video captioning model capable of both utterance transcription and speaker attribution, while preserving holistic audiovisual context. Building upon AVoCaDO, we first design a dedicated pipeline to synthesize high-quality dialogue-aware audiovisual captions for SFT. We then incorporate human-annotated samples to perform a difficulty-partitioned two-stage RL strategy to further enhance dialogue description quality. Experimental results demonstrate that DiaDem not only outperforms the Gemini series in dialogue description accuracy, but also achieves competitive performance on general audiovisual captioning benchmarks. Comprehensive ablation studies further confirm the contribution of each component in our training pipeline, underscoring the effectiveness of the proposed approach.

## Acknowledgements

This work is supported by Beijing Natural Science Foundation (L252033) and National Natural Science Foundation of China (92570204, 62576339).

## References


1. Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Deep audio-visual speech recognition. IEEE transactions on pattern analysis and machine intelligence **44**(12), 8717–8727 (2018) 4
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 13
3. Bellman, R.: Dynamic programming. science **153**(3731), 34–37 (1966) 5
4. Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: Noise reduction in speech processing, pp. 1–4. Springer (2009) 13
5. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska Masson, A., Mccowan, I., Post, W., Reidsma, D., Wellner, P.: The ami meeting corpus: A pre-announcement. In: Lecture Notes in Computer Science (07 2005). https://doi.org/10.1007/11677482_3 4
6. Chai, W., Song, E., Du, Y., Meng, C., Madhavan, V., Bar-Tal, O., Hwang, J.N., Xie, S., Manning, C.D.: Auroracap: Efficient, performant video detailed captioning and a new benchmark. arXiv preprint arXiv:2410.03051 (2024) 1
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) 2, 3, 10
8. Chen, Y., Zheng, S., Wang, H., Cheng, L., Zhu, T., Huang, R., Deng, C., Chen, Q., Zhang, S., Wang, W., et al.: 3d-speaker-toolkit: An open-source toolkit for multi-modal speaker verification and diarization. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025) 4
9. Cheng, M., Li, M.: Multi-input multi-output target-speaker voice activity detection for unified, flexible, and robust audio-visual speaker diarization. IEEE Transactions on Audio, Speech and Language Processing (2025) 4
10. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024) 3
11. Chung, J.S., Huh, J., Nagrani, A., Afouras, T., Zisserman, A.: Spot the conversation: speaker diarisation in the wild. In: Interspeech (2020) 4
12. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025) 3, 10
13. Cornell, S., Jung, J.w., Watanabe, S., Squartini, S.: One model to rule them all? towards end-to-end joint speaker diarization and speech recognition. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 11856–11860. IEEE (2024) 4

14. Cornell, S., Wiesner, M., Watanabe, S., Raj, D., Chang, X., Garcia, P., Maciejewski, M., Masuyama, Y., Wang, Z.Q., Squartini, S., et al.: The chime-7 dasr challenge: Distant meeting transcription with multiple devices in diverse scenarios. arXiv preprint arXiv:2306.13734 (2023) 4
15. 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) 2
16. Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Transactions on Multimedia **2**(3), 141–151 (2000). https://doi.org/10.1109/6046.865479 4
17. Efstathiadis, G., Yadav, V., Abbas, A.: Llm-based speaker diarization correction: A generalizable approach. Speech Communication **170**, 103224 (2025) 4
18. 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) 10
19. Guo, Y., Ma, S., Ma, S., Bao, X., Xie, C.W., Zheng, K., Weng, T., Sun, S., Zheng, Y., Zou, W.: Aligned better, listen better for audio-visual large language models. arXiv preprint arXiv:2504.02061 (2025) 3
20. Hou, W., Li, G., Tian, Y., Hu, D.: Toward long form audio-visual video understanding. ACM Transactions on Multimedia Computing, Communications and Applications **20**(9), 1–26 (2024) 3
21. Kanda, N., Gaur, Y., Wang, X., Meng, Z., Chen, Z., Zhou, T., Yoshioka, T.: Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers. arXiv preprint arXiv:2006.10930 (2020) 4
22. Liang, Y., Shi, M., Yu, F., Li, Y., Zhang, S., Du, Z., Chen, Q., Xie, L., Qian, Y., Wu, J., et al.: The second multi-channel multi-party meeting transcription challenge (m2met 2.0): A benchmark for speaker-attributed asr. In: 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 1–8. IEEE (2023) 4
23. Liu, T., Fan, S., Xiang, X., Song, H., Lin, S., Sun, J., Han, T., Chen, S., Yao, B., Liu, S., Wu, Y., Qian, Y., Yu, K.: MSDWild: Multi-modal Speaker Diarization Dataset in the Wild. In: Proc. Interspeech 2022. pp. 1476–1480 (2022). https://doi.org/10.21437/Interspeech.2022-10466 4
24. Lu, L., Kanda, N., Li, J., Gong, Y.: Streaming multi-talker speech recognition with joint speaker identification. arXiv preprint arXiv:2104.02109 (2021) 4
25. 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) 2, 3, 10
26. Medennikov, I., Korenevsky, M., Prisyach, T., Khokhlov, Y., Korenevskaya, M., Sorokin, I., Timofeeva, T., Mitrofanov, A., Andrusenko, A., Podluzhny, I., et al.: Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario. arXiv preprint arXiv:2005.07272 (2020) 4
27. Mroueh, Y., Marcheret, E., Goel, V.: Deep multimodal learning for audio-visual speech recognition. In: 2015 IEEE International conference on acoustics, speech and signal processing (ICASSP). pp. 2130–2134. IEEE (2015) 4
28. von Neumann, T., Boeddeker, C., Delcroix, M., Haeb-Umbach, R.: Meeteval: A toolkit for computation of word error rates for meeting transcription systems. arXiv preprint arXiv:2307.11394 (2023) 7

29. von Neumann, T., Boeddeker, C., Delcroix, M., Haeb-Umbach, R.: Word error rate definitions and algorithms for long-form multi-talker speech recognition. IEEE Transactions on Audio, Speech and Language Processing (2025) 7
30. von Neumann, T., Boeddeker, C., Kinoshita, K., Delcroix, M., Haeb-Umbach, R.: On word error rate definitions and their efficient computation for multi-speaker speech recognition systems. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023) 7
31. Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G., Ogata, T.: Audio-visual speech recognition using deep learning. In: Applied Intelligence. pp. 722–737 (2015) 4
32. Panagopoulou, A., Xue, L., Yu, N., Li, J., Li, D., Joty, S., Xu, R., Savarese, S., Xiong, C., Niebles, J.C.: X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. arXiv preprint arXiv:2311.18799 (2023) 3
33. Park, T.J., Dhawan, K., Koluguri, N., Balam, J.: Enhancing speaker diarization with large language models: A contextual beam search approach. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 10861–10865. IEEE (2024) 4
34. Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., Pantic, M.: End-to-end audiovisual speech recognition. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 6548–6552. IEEE (2018) 4
35. Raj, D., Denisov, P., Chen, Z., Erdogan, H., Huang, Z., He, M., Watanabe, S., Du, J., Yoshioka, T., Luo, Y., et al.: Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis. In: 2021 IEEE spoken language technology workshop (SLT). pp. 897–904. IEEE (2021) 4
36. Ren, Y., Lin, Z., Li, Y., Meng, G., Wang, W., Wang, J., Lin, Z., Dai, J., Yang, Y., Wang, W., et al.: Anycap project: A unified framework, dataset, and benchmark for controllable omni-modal captioning. arXiv preprint arXiv:2507.12841 (2025) 2
37. Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., Ramaswamy, S., Stopczynski, A., Schmid, C., Xi, Z., et al.: Ava active speaker: An audio-visual dataset for active speaker detection. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 4492–4496. IEEE (2020) 4
38. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017) 9
39. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024) 9
40. Shu, F., Zhang, L., Jiang, H., Xie, C.: Audio-visual llm for video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4246–4255 (2025) 3
41. Sklyar, I., Piunova, A., Zheng, X., Liu, Y.: Multi-turn rnn-t for streaming recognition of multi-party speech. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8402–8406. IEEE (2022) 7
42. Sun, G., Yu, W., Tang, C., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., Wang, Y., Zhang, C.: video-salmonn: Speech-enhanced audio-visual large language models. arXiv preprint arXiv:2406.15704 (2024) 3

43. Tang, C., Li, Y., Yang, Y., Zhuang, J., Sun, G., Li, W., Ma, Z., Zhang, C.: video-salmonn 2: Captioning-enhanced audio-visual large language models. arXiv preprint arXiv:2506.15220 (2025) 1, 2, 3, 10
44. Team, M.L., Wang, B., Xiao, B., Zhang, B., Rong, B., Chen, B., Wan, C., Zhang, C., Huang, C., Chen, C., et al.: Longcat-flash-omni technical report. arXiv preprint arXiv:2511.00279 (2025) 3
45. Wang, J., Yuan, L., Zhang, Y., Sun, H.: Tarsier: Recipes for training and evaluating large video description models. arXiv preprint arXiv:2407.00634 (2024) 1
46. Wang, Q., Huang, Y., Zhao, G., Clark, E., Xia, W., Liao, H.: Diarizationlm: Speaker diarization post-processing with large language models. arXiv preprint arXiv:2401.03506 (2024) 4
47. Wang, W., Cai, D., Cheng, M., Li, M.: Joint inference of speaker diarization and asr with multi-stage information sharing. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 11011–11015 (2024). https://doi.org/10.1109/ICASSP48485.2024.10446724 4
48. 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. arXiv preprint arXiv:2502.20811 (2025) 2
49. Watanabe, S., Mandel, M., Barker, J., Vincent, E., Arora, A., Chang, X., Khudanpur, S., Manohar, V., Povey, D., Raj, D., et al.: Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. arXiv preprint arXiv:2004.09249 (2020) 7
50. 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) 1, 2, 3, 10
51. Wuerkaixi, A., Yan, K., Zhang, Y., Duan, Z., Zhang, C.: Dyvise: Dynamic vision-guided speaker embedding for audio-visual speaker diarization. In: 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP). pp. 1–6 (2022). https://doi.org/10.1109/MMSP55362.2022.9948860 4
52. Xu, E.Z., Song, Z., Tsutsui, S., Feng, C., Ye, M., Shou, M.Z.: Ava-avd: Audio-visual speaker diarization in the wild. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3838–3847 (2022) 4
53. Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., et al.: Qwen2. 5-omni technical report. arXiv preprint arXiv:2503.20215 (2025) 10
54. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., Lv, Y., Wang, Y., Guo, D., Wang, H., Ma, L., Zhang, P., Zhang, X., Hao, H., Guo, Z., Yang, B., Zhang, B., Ma, Z., Wei, X., Bai, S., Chen, K., Liu, X., Wang, P., Yang, M., Liu, D., Ren, X., Zheng, B., Men, R., Zhou, F., Yu, B., Yang, J., Yu, L., Zhou, J., Lin, J.: Qwen3-omni technical report. arXiv preprint arXiv:2509.17765 (2025) 10
55. 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) 10
56. Ye, H., Yang, C.H.H., Goel, A., Huang, W., Zhu, L., Su, Y., Lin, S., Cheng, A.C., Wan, Z., Tian, J., et al.: Omnivinci: Enhancing architecture and data for omni-modal understanding llm. arXiv preprint arXiv:2510.15870 (2025) 10
57. Ye, Q., Yu, Z., Shao, R., Xie, X., Torr, P., Cao, X.: Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In: European Conference on Computer Vision. pp. 146–164. Springer (2024) 3

58. Yin, H., Chen, Y., Deng, C., Cheng, L., Wang, H., Tan, C.H., Chen, Q., Wang, W., Li, X.: Speakerlm: End-to-end versatile speaker diarization and recognition with multimodal large language models. arXiv preprint arXiv:2508.06372 (2025) 4
59. 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) 1
60. Zhang, Y., Li, Z., Wang, D., Zhang, J., Zhou, D., Yin, Z., Dai, X., Yu, G., Li, X.: Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. arXiv preprint arXiv:2507.09862 (2025) 2
61. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024) 2
62. Zheng, X., Zhang, C., Woodland, P.: Dncasr: End-to-end training for speaker-attributed asr. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 18369–18383 (2025) 4
63. Zheng, X., Zhang, C., Woodland, P.C.: Tandem multitask training of speaker diarisation and speech recognition for meeting transcription. arXiv preprint arXiv:2207.03852 (2022) 4
64. 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) 1