D-ORCA: DIALOGUE-CENTRIC OPTIMIZATION FOR ROBUST AUDIO-VISUAL CAPTIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spoken dialogue is a primary source of information in videos; therefore, accurately identifying who spoke what and when is essential for deep video understanding. We introduce D-ORCA, a **d**ialogue-centric **o**mni-modal large language model optimized for **r**obust audio-visual **ca**ptioning. We further curate DVD, a large-scale, high-quality bilingual dataset comprising nearly 40,000 multi-party dialogue videos for training and 2000 videos for evaluation in English and Mandarin, addressing a critical gap in the open-source ecosystem. To ensure fine-grained captioning accuracy, we adopt group relative policy optimization with three novel reward functions that assess speaker attribution accuracy, global speech content accuracy, and sentence-level temporal boundary alignment. These rewards are derived from evaluation metrics widely used in speech processing and, to our knowledge, are applied for the first time as reinforcement learning objectives for audio-visual captioning. Extensive experiments demonstrate that D-ORCA substantially outperforms existing open-source models in speaker identification, speech recognition, and temporal grounding. Notably, despite having only 8 billion parameters, D-ORCA achieves performance competitive with Qwen3-Omni across several general-purpose audio-visual understanding benchmarks. Demos are available at https://d-orca-llm.github.io/, and our code, data, and checkpoints will be released.

1 INTRODUCTION

Dialogue-centric content (*e.g.*, films, television dramas, and everyday vlogs) constitutes a substantial portion of modern video data. In these scenarios, spoken dialogue drives the narrative, requiring audio-visual understanding systems to resolve not only what is said, but also who says it and when. Although recent visual-only (Zhang et al., 2024; Li et al., 2024; Liu et al., 2024; Zhang et al., 2025; Chai et al., 2025; Zhu et al., 2025; Li et al., 2025; Bai et al., 2025b;a; Yao et al., 2024) and audio–visual (Comanici et al., 2025; Xu et al., 2025a; Tang et al., 2025; Ma et al., 2025; Chen et al., 2025; Ge et al., 2025; Sun et al., 2025; Ye et al., 2025) large language models (LLMs) offer promising end-to-end solutions, a critical gap remains: unlike humans, who naturally integrate auditory and visual cues (*e.g.*, synchronizing voice with lip movements, or inferring off-screen speakers from static lips) for speaker attribution, current audio-visual LLMs struggle to achieve this level of cross-modal integration. As shown in Fig. 1, most existing open-source models fail to produce accurate, dialogue-aware captions, exposing a key limitation in current multimodal architectures.

The field also lacks benchmarks and evaluation metrics tailored to dialogue-centric audio–visual captioning. Standard captioning metrics such as BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) are inadequate for long-form dialogue, as they fail to capture semantic fidelity and speaker-specific accuracy. Recent metrics for detailed video captioning (Chai et al., 2025; Tang et al., 2025; Ma et al., 2025) emphasize overall content coverage but overlook the evaluation of synchronous audio–visual interactions essential to dialogue-driven scenarios. As a result, existing open-source models lack effective optimization signals for dialogue-centric understanding, often failing to correctly attribute speech to speakers or maintain coherent dialogue narratives.

To address these challenges, we first tackle the data scarcity problem by curating a large-scale video training set, DVD-Train, in dialogue-centric scenarios annotated with fine-grained **d**ialogue-centric **v**ideo **d**escriptions. Building on this foundation, we further construct DVD-Bench, a high-quality video benchmark for dialogue-centric audio-visual understanding, which comprises diverse, human-

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



Annotation:
0:01 - 0:05, Judy Hopps (the rabbit): Sir, you were going 115 miles per hour. I hope you have a good explanation.
0:06 - 0:09, Nick Wilde (the fox): Flash, Flash, Hundred Yard Dash!
0:13 - 0:15, Flash (the sloth): Nick.

Qwen3-Omni-Flash:
Judy Hopps (voice): "Sir, you were going 115 miles per hour. I hope you have a good explanation." ...
Flash (voice): "Flash. Flash. Hundred yard dash." ...
Flash (cont'd): "Nick." ...
Nick Wilde (voice, off-screen or implied): He doesn't speak in this clip, ...

ARC-Hunyuan-Video-7B:
00:00:00 - 00:00:05 A rabbit police officer pulls over a sloth for speeding and demands an explanation.
00:00:05 - 00:00:10 The sloth, Nick, confidently yells 'hundred yard dash' as a defense.
00:00:10 - 00:00:15 Realizing he is in trouble, the sloth begins to stammer and awkwardly try to name himself, 'Nick.'

AVoCaDO:
... A female voice speaks "Sir, you were going 115 miles per hour. I hope you have a good explanation."...
A male voice responds, sounding defensive and a bit flustered as he tries to justify his actions. "Flash, flash, hundred-yard dash." After a brief pause, his tone shifts to one of proud, almost smug satisfaction as he concludes, "Nick."

D-ORCA:
0:01 - 0:05, Judy Hopps said: "Sir, you were going 115 miles per hour. I hope you have a good explanation." ...
0:06 - 0:09, Nick Wilde said: "Flash, flash, hundred yard dash." ...
From 0:09 to 0:12, the sloth stares at Nick and Judy, processing what they said. Then, the sloth begins to speak.
0:13 - 0:15, the sloth said: "Nick." ...

Figure 1: An example of video captions generated by different audio-visual LLMs for a dialogue-centric video clip. Green highlights indicate correct speaker attribution, red indicates incorrect attribution, and orange marks ambiguous or implicit speaker references. Existing open-source audio-visual LLMs struggle to accurately comprehend video dialogue, while our D-ORCA demonstrates more accurate and robust audio-visual understanding for dialogue-centric scenarios.

annotated English and Chinese dialogue scenes to enable rigorous and systematic evaluation. We argue that the core challenge of dialogue-centric video captioning lies in resolving the triplet of "**when**, **who**, and **what** is said". This requires accurately localizing utterances in time, associating them with the correct speakers, and precisely recognizing the spoken content. Despite recent advances, these abilities remain challenging for modern artificial intelligence (AI) models, and thus naturally define critical evaluation dimensions for coherent, dialogue-driven audio-visual understanding. To explicitly optimize audio-visual LLMs for these capabilities, we propose a novel reinforcement learning (RL) framework that integrates group relative policy optimization (GRPO) (Shao et al., 2024) with three specialized reward functions, which guide the model toward fine-grained audio–visual alignment tailored to dialogue-centric scenarios:

- A speaker attribution accuracy reward that improves the binding between spoken content and the correct speaker;
- A global speech content reward that enhances automatic speech recognition (ASR) accuracy;
- A sentence-level temporal boundary reward that encourages precise utterance boundary localization.

Notably, these reward functions also serve as effective evaluation metrics for dialogue-centric audio-visual captioning. To ensure training stability and robustness, we introduce a pre–direct preference optimization (DPO) warm-up stage along with a reward balancing mechanism. Experiments show that the resulting model with 8 billion (B) parameters, D-ORCA, achieves state-of-the-art (SOTA) performance among open-source models for dialogue-centric audio-visual captioning in both English and Chinese. D-ORCA shows superior capability in speaker identification, precise temporal grounding, and robust speech recognition. Moreover, it exhibits strong generalization, achieving SOTA performance on general audio-visual question answering (QA) benchmarks among models of comparable scale. Our main contributions are summarized as follows:

- We curate a large-scale, dialogue-rich bilingual video dataset DVD-Train as well as a high-quality benchmark DVD-Bench to advance research in dialogue-centric audio-visual understanding.
- We propose a robust RL-based post-training framework with novel reward signals explicitly designed to resolve the challenges of "when, who, and what is said".

- We present D-ORCA, which achieves SOTA results in dialogue-centric captioning while remaining highly competitive on broad audio-visual QA benchmarks.

2 BACKGROUND

2.1 DIALOGUE-CENTRIC AUDIO-VISUAL UNDERSTANDING

The core of dialogue-centric audio-visual understanding lies in resolving the triplet of "when, who, and what is said". Traditionally, the audio-only formulation of this problem has been investigated through speaker diarization pipelines, which typically involve voice activity detection to identify speech segments, speaker change point detection to partition them into speaker-homogeneous segments, and speaker embedding extraction followed by clustering for speaker identity assignment (Park et al., 2022). To recover linguistic content, these modules are typically integrated into cascaded speaker-attributed ASR (SA-ASR) systems, where diarized segments are transcribed by downstream ASR models (Raj et al., 2021; Zheng et al., 2022; Cornell et al., 2023). To reduce error propagation, recent work has shifted toward end-to-end SA-ASR. Existing approaches either assume a fixed speaker set and use multi-head architectures for speaker-wise transcription (Kanda et al., 2020; Lu et al., 2021), or handle open speaker scenarios via neural clustering or speaker indexing (Zheng et al., 2025). More recently, the emergence of LLMs has introduced a new paradigm. Generative approaches such as SpeakerLM (Yin et al., 2025) and DiarizationLM (Wang et al., 2024) have shown the potential of leveraging LLMs to perform end-to-end diarization and recognition within a unified sequence generation task.

In audio–visual settings, acoustic information alone is often insufficient, yet extending audio-centric pipelines to effectively incorporate visual cues remains challenging. Conventional approaches lack principled mechanisms to extract heterogeneous visual signals needed for accurate speaker attribution. Prior methods that rely on facial dynamics for diarization (Xu et al., 2022; Sharma & Narayanan, 2022; Wuerkaixi et al., 2022; He et al., 2022; Qiu et al., 2022; Yang et al., 2023; Cheng & Li, 2025; Afouras et al., 2018) are brittle in complex scenarios (e.g., off-screen speakers) and lack holistic video understanding. The emergence of audio–visual LLMs that jointly process raw audio and video offers a promising alternative. A representative example is AVoCaDO (Chen et al., 2025), which introduces dialogue-based rewards to optimize captioning based on speech content and speaker identity.

Traditional video captioning benchmarks such as MSR-VTT (Xu et al., 2016) and VATEX (Wang et al., 2019) primarily rely on short-form descriptions evaluated using n-gram–based metrics (*e.g.*, BLEU, ROUGE-L), which are not suitable for assessing the fine-grained semantics of complex audio-visual scenes. As a result, recent work has increasingly focused on evaluation and optimization for detailed video captioning. In the visual domain, AuroraCap (Chai et al., 2025) introduced a QA-based evaluation paradigm to measure content coverage. This idea was later extended to the audio-visual setting by UGC-VideoCaptioner (Wu et al., 2025), which employs dual-modality question answering for evaluation and adopts GRPO with LLM-scored rewards during training. Similarly, AVoCaDO (Chen et al., 2025) leverages GRPO to refine captioning quality using checklist-based and dialogue-based rewards verified by external LLMs. video-SALMONN 2 (Tang et al., 2025) achieves fine-grained assessment by decomposing video content into atomic events to quantify completeness versus hallucination, and guides model learning through multi-round DPO. Omni-Captioner (Ma et al., 2025) emphasizes data scaling via an agentic framework that synthesizes dense annotations through iterative query–observation cycles. Despite these advances, existing approaches for detailed audio-visual captioning largely emphasize the completeness of auditory and visual content, while overlooking the quality of joint audio-visual interactions, particularly those required for dialogue-centric captioning, such as speaker attribution, temporal alignment, and cross-modal consistency.

3 METHODS

3.1 MODEL ARCHITECTURE AND TRAINING PIPELINE

Since current omni-modal LLMs either lack strong visual performance or require excessive audio tokens (e.g., Qwen2.5-Omni at 25 tokens/second), D-ORCA is not built upon existing omni-modal LLMs. Instead, it integrates audio capabilities into a powerful visual LLM (e.g., Qwen3-VL) while maintaining a low audio token rate (e.g., 2 tokens/second) to accelerate both training and inference.

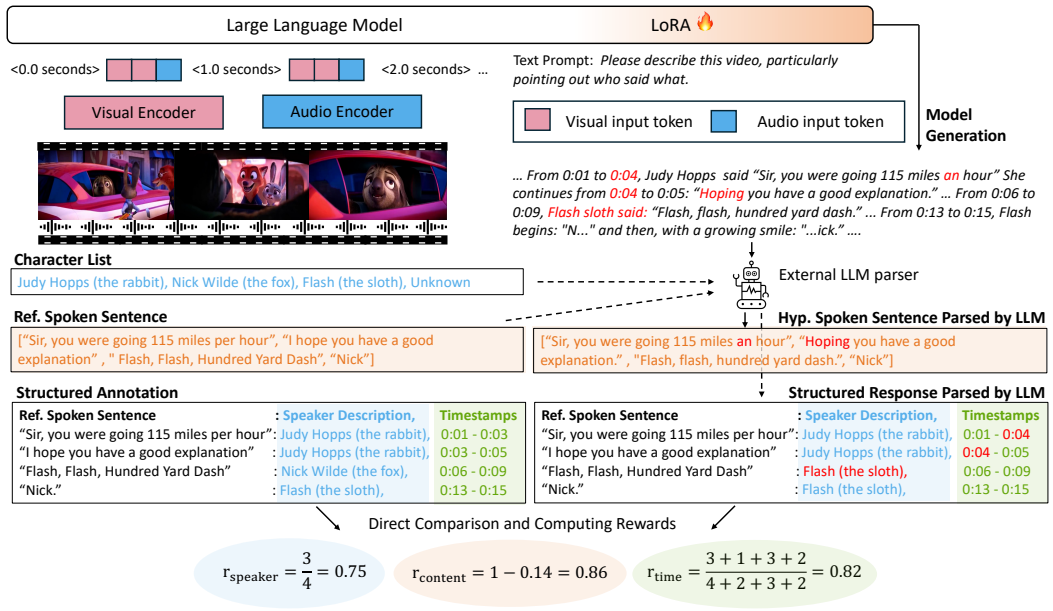


Figure 2: Architecture and reward computation in D-ORCA. The model processes chronologically interleaved audio–visual tokens as input. During GRPO training, an external LLM parser is used to extract predicted ASR texts, and also identify the speaker and timestamp for each reference spoken sentence guided by a candidate character list. Based on these structured outputs, dialogue-centric rewards (r_{speaker}, r_{content}, r_{time}) are computed to optimize the accuracy of speaker attribution, speech content, and sentence-level temporal alignment.

As illustrated in Fig. 2, our architecture is designed to process paired audio–visual streams conditioned on user text prompts. The pipeline begins with parallel feature extraction from the input video. The visual stream is first temporally down-sampled into a sequence of frames and processed frame by frame using a pre-trained visual encoder. In parallel, the accompanying audio stream is processed by a pre-trained audio encoder. Both encoders are equipped with modality aligners that project the extracted visual and acoustic features into visual and audio input tokens compatible with the LLM. These modality-specific tokens are grouped according to their corresponding time intervals and interleaved to form a unified temporal sequence. Explicit textual timestamps are inserted as delimiters at the beginning of each temporal segment. The resulting multimodal token sequence is then concatenated with the user's text prompt and fed into the LLM backbone. To improve training efficiency and parameter economy, we apply low-rank adaptation (LoRA) (Hu et al., 2022) to the LLM.

As for the training pipeline, the model is built upon a strong visual LLM. To endow this visual backbone with auditory perception and dialogue understanding capabilities, we first conduct audio modality alignment using large-scale audio datasets, followed by audio-visual supervised fine-tuning (SFT) on paired video-text data. Specifically, in these stages, given the text prompt $\mathbf{q}$, the multimodal input sequence $\mathbf{x}$, and the target textual response $\mathbf{y}$, the training objective is to minimize the negative log-likelihood of the next-token prediction:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(\mathbf{x},\mathbf{y},\mathbf{q})\sim\mathcal{D}}\left[\log \pi_\theta(\mathbf{y}|\mathbf{x},\mathbf{q})\right], \tag{1}$$

where $\mathcal{D}$ and θ are the training set and model parameters.

Prior to optimizing dialogue-centric captioning with GRPO, we introduce a pre-DPO stage to mitigate the SFT model's repetition degeneration, which would otherwise destabilize reward calculation and hinder effective optimization. In the pre-DPO stage, $\forall \mathbf{x} \in \mathcal{D}$, two candidate captions are sampled from the SFT model. Inputs for which exactly one caption is complete and the other exhibits repetitive degeneration are retained to construct the pre-DPO training set $\mathcal{D}_{\text{DPO}}$. The complete caption and the repetitive caption are then assigned as the positive and negative samples, denoted by $\mathbf{y}_{\text{win}}$ and $\mathbf{y}_{\text{lose}}$, respectively. Under this construction, the training loss of the pre-DPO stage is defined as:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_{\text{win}},\mathbf{y}_{\text{lose}})\sim\mathcal{D}_{\text{DPO}}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(\mathbf{y}_{\text{win}}\mid\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{win}}\mid\mathbf{x})} - \beta\log\frac{\pi_\theta(\mathbf{y}_{\text{lose}}\mid\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_{\text{lose}}\mid\mathbf{x})}\right)\right], \tag{2}$$

where β controls the deviation from the reference model π_{ref}, and $\sigma(\cdot)$ is the sigmoid function.

Finally, GRPO is applied to optimize dialogue-centric video captioning, and a reward computation example is also shown in Fig 2. Specifcally, for each input prompt $\mathbf{q}$, a group of G outputs $\{\mathbf{y}_g\}_{g=1}^{G}$ from the old policy $\pi_{\theta_{\text{old}}}$ are generated, and we calculate rewards and normalize them within the group to obtain advantages A_g for each output. The objective loss is formulated to maximize the expected reward while constraining policy updates, as shown below:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{(\mathbf{x},\mathbf{q})\sim\mathcal{D},\{\mathbf{y}_g\}\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{G}\sum_{g=1}^{G}\min\left(\rho_g A_g, \text{clip}(\rho_g, 1-\epsilon, 1+\epsilon)A_g\right)\right], \tag{3}$$

where $\rho_g = \pi_\theta(\mathbf{y}_g|\mathbf{x},\mathbf{q})/\pi_{\theta_{\text{old}}}(\mathbf{y}_g|\mathbf{x},\mathbf{q})$ is the probability ratio, and ϵ is a hyper parameter of clip$(\cdot)$ that limits $1-\epsilon \leqslant \rho_g < 1+\epsilon$. Note that the KL-divergence term in standard GRPO (Shao et al., 2024) is not used.

3.2 Dialogue-centric Reward Modeling

To optimize the model's ability to resolve when, who, and what is spoken in video dialogues, we propose a fine-grained reward modeling framework. Specifically, we design three complementary reward signals: *Speaker Attribution Accuracy* (r_{speaker}), *Global Speech Content Accuracy* (r_{content}), and *Sentence-level Temporal Boundary Alignment* (r_{time}). An advanced external LLM is employed to parse the generated captions into a structured form, enabling direct comparison with the ground-truth annotations.

Data Structure and Annotation. We formally represent the annotation for a video as a set of spoken sentence annotations $\mathcal{V} = (s_i, p_i, \tau_i)_{i=1}^{N}$, where s_i denotes the reference transcript of the i-th utterance, p_i specifies the speaker identity via a textual description, and $\tau_i = [t_{i,\text{start}}, t_{i,\text{end}}]$ defines the corresponding temporal interval. Here, N denotes the total number of spoken sentences in the video. In addition, we maintain a global set of character descriptions $\mathcal{C}$ that covers all distinct roles appearing in the video. Note that some characters in $\mathcal{C}$ may not speak, and speaking characters may not appear within the video frames.

Speaker Attribution Accuracy Reward r_{speaker}. This reward measures the model's ability to correctly bind speech content to the corresponding speakers. We employ an external LLM to parse the generated caption and assign each ground-truth sentence s_i to a predicted speaker $\hat{p}_i$ selected from a candidate set $\mathcal{C}$. To penalize ambiguous speaker attribution, we augment $\mathcal{C}$ with an "Unknown" option during training; sentences with unspecified or unresolvable speakers are assigned to this class. The reward is computed as the proportion of sentences with correctly attributed speakers:

$$r_{\text{speaker}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{p}_i = p_i), \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Global Speech Content Reward r_{content}. To ensure high-quality speech recognition within the captioning process, we evaluate transcription accuracy using word error rate (WER) for English and character error rate (CER) for Chinese. To reduce errors caused by cross-sentence word or character misalignment, we concatenate all ground-truth utterances $\mathbf{s}_i$ and the corresponding hypothesis utterances $\hat{s}_i$, extracted in chronological order from the generated captions by an external LLM, into unified transcripts S and $\hat{S}$ for each video $\mathcal{V}$. The resulting content reward is then defined as:

$$r_{\text{content}} = \begin{cases} 1 - \text{WER}(S,\hat{S}), & \text{for English,} \\ 1 - \text{CER}(S,\hat{S}), & \text{for Chinese.} \end{cases} \tag{5}$$

Sentence-level Temporal Reward r_{time}. This reward encourages precise temporal localization of dialogue events. An external LLM extracts predicted start and end timestamps $[\hat{t}_{i,\text{start}}, \hat{t}_{i,\text{end}}]$ for each sentence s_i from the generated caption, and rewards are computed based on intersection over

union (IoU). We explicitly handle two unknown (Unk) cases during LLM parsing: **(i) Bilateral Unknown**, where no timestamp is produced and the predicted interval is set to [Unk, Unk], yielding zero intersection and a union equal to the ground-truth duration $|\tau_i|$; and **(ii) Unilateral Unknown**, where only one boundary is generated (*e.g.*, due to partial timestamps or sentence misalignment; see Appendix B). In the latter case, instead of assigning zero IoU, we estimate the missing boundary by projecting the ground-truth duration $|\tau_i|$ (*e.g.*, $[\hat{t}_{i,\text{start}}, \text{Unk}] \rightarrow [\hat{t}_{i,\text{start}}, \hat{t}_{i,\text{start}} + |\tau_i|]$).

After resolving Unk cases, we compute the intersection I_i and union U_i between predicted and ground-truth intervals. The reward is defined as the global IoU across all sentences:

$$r_{\text{time}} = \frac{\sum_{i=1}^{N} I_i}{\sum_{i=1}^{N} U_i}. \tag{6}$$

Curriculum Learning Strategy. The final reward r_{total} is computed as a weighted sum of the three components. To explicitly penalize repetition degeneration, we impose a strict length constraint: if the generated response contains more than l_{max} tokens, r_{total} is set to zero. In addition, we observe that applying temporal constraints too early in training can cause optimization collapse. To address this issue, we adopt a curriculum strategy in which the temporal reward is introduced only after the model has sufficiently stabilized. The total reward at training step k is defined as:

$$r_{\text{total}} = \begin{cases} 0 & \text{if } l > l_{\text{max}}, \\ \lambda_1 r_{\text{speaker}} + \lambda_2 r_{\text{content}} & \text{else if } k < K_{\text{warmup}}, \\ \lambda_1 r_{\text{speaker}} + \lambda_2 r_{\text{content}} + \lambda_3 r_{\text{time}} & \text{otherwise.} \end{cases} \tag{7}$$

where $\lambda_1, \lambda_2, \lambda_3$ are balancing coefficients and K_{warmup} is the threshold step for curriculum activation.

4 Experimental Setup

4.1 Data Specifications

For audio modality alignment, we train on LibriSpeech (Panayotov et al., 2015), Common-Voice (Ardila et al., 2020), WavCaps (Mei et al., 2024), AudioCaps (Kim et al., 2019), Clotho (Drossos et al., 2020), and a 3,500-hour (h) subset of WenetSpeech (Zhang et al., 2022).

For the subsequent audio-visual training stages, we construct a large-scale, high-quality video dataset, DVD-Train, explicitly focused on dialogue-rich scenarios. The videos are manually sourced from diverse genres, including movie clips, TV dramas, TV shows, and interviews, ensuring a wide coverage of speaking styles and acoustic environments. Specifically, we compile a total of about 40,000 videos, evenly split between English and Chinese, to serve as the foundation for the audio-visual SFT and pre-DPO. To support the reward-based optimization, we further curate a specific subset from this primary collection, consisting of approximately 8,500 videos randomly selected from DVD-Train, with a roughly equal distribution of English and Chinese content.

For evaluation, we construct DVD-Bench, comprising 964 English and 1,014 Chinese videos with complex multi-party dialogue. Unlike the training data, DVD-Bench is carefully curated and annotated by human experts who provide precise speaker identities for each spoken sentence, establishing a reliable benchmark for evaluation. Note that when testing on DVD-Bench, the model can only rely on the raw video to output captions, and cannot be given other information like the candidate character list. Details of the DVD dataset are provided in Appendix C.

4.2 Model, training and evaluation specifications

The visual encoder and the backbone LLM of D-ORCA are initialized from Qwen3-VL-8B (Bai et al., 2025a), while the audio encoder utilizes Whisper-large-v3 (Radford et al., 2023). To map the Whisper features into the LLM's input space, a randomly initialized window-level Q-Former (Tang et al., 2024) is used as the audio modality aligner. This Q-Former is configured with a single query token and operates on a temporal window of 0.5 seconds, effectively compressing the audio input into a sequence of one token per 0.5 seconds. For visual processing, inputs are sampled at a frame rate of 2 frames per second and capped at a maximum of 128 frames, with a resolution limit of 176,400 pixels per frame. This results in a maximum number of visual tokens of about 10,000. The LoRA of the LLM is set to a rank of 128 and a scaling factor of 2.

In the initial audio modality alignment stage, only the audio modality aligner is trained while the LLM remains frozen. During subsequent audio–visual SFT, both modality aligners and LoRA parameters

are unfrozen to enable joint multimodal learning. In the pre-DPO and GRPO stages, optimization is restricted to LoRA parameters. LoRA merging is performed during pre-DPO following video-SALMONN 2 (Tang et al., 2025). For GRPO, we set the group size to 8. Hyperparameters in Equ. (7) are configured as L_{max}=3072, K_{warmup}=500, λ_1=0.9, λ_2=0.1, and λ_3=0.1, with a learning rate of 2×10^{-5}. Training cost details are reported in Appendix A.

To evaluate dialogue-centric audio–visual captioning, we mirror the reward design and measure speaker attribution accuracy, speech transcription fidelity in WER/CER, and sentence-level temporal accuracy in IoU. Consistent with GRPO training, we use Gemini-2.5-Flash as the external judge for caption parsing and evaluation.

5 Experimental Results

5.1 Main Results

Table 1 presents a comprehensive quantitative comparison between D-ORCA and SOTA open-source omni-modal LLMs on our dialogue-centric audio-visual captioning benchmark. We also provide an audio-only Speaker-Attributed ASR (SA-ASR) baseline based on WhisperX (Bain et al., 2023). Since the SA-ASR pipeline lacks visual perception, it cannot directly map speaker IDs to specific characters. Therefore, for each sample, we exhaustively evaluate all possible mappings between speaker IDs and the character candidate set $\mathcal{C}$, and report the highest speaker attribution accuracy.

The results indicate that D-ORCA provides a holistic and structured understanding of the video narrative, demonstrating a distinct advantage in resolving the "when, who, what is said" of audio-visual interactions. On the primary metric of speaker attribution accuracy, D-ORCA achieves strong performance in both English and Chinese settings. Notably, our base SFT model, trained directly via standard next-token prediction, achieves a moderate accuracy of 71.2% (En) and 69.6% (Zh). While this surpasses most open-source omni-modal baselines and matches AVoCaDO, it still falls significantly short of powerful closed-source models like Gemini-3.0-flash. This substantial performance gap highlights the fundamental limitation of relying solely on SFT for complex dialogue-centric reasoning.

Building on this, our proposed RL method, utilizing only about 8,500 samples for training, delivers significant additional gains. The final D-ORCA model improves speaker attribution accuracy by approximately 10% over the SFT baseline. This leap allows our 8B model to effectively match the performance of Gemini-3.0-flash, firmly validating the efficacy and necessity of our dialogue-centric optimization algorithm over standard supervised learning.

Besides, most existing audio-visual LLMs, such as ARC-Qwen-Video-Narrator, video-SALMONN 2+, Qwen2.5-Omni, and Qwen3-Omni, tend to generate coarse dialogue summaries, often omitting key utterances or failing to capture exact speech content, which leads to higher word error rates. In contrast, D-ORCA achieves consistently low WER/CER, matching the audio-only SA-ASR baseline in English and surpassing it in Chinese, while approaching the transcription quality of Gemini-3.0-flash. This capability is initially established during the SFT stage through verbatim-oriented supervision and is further enhanced by RL.

With respect to temporal precision, D-ORCA achieves fine-grained temporal awareness. Through RL training, D-ORCA substantially improves its temporal localization performance, achieving IoU scores of 57.1% and 37.8% on English and Chinese videos, respectively. This demonstrates a massive improvement over its SFT baseline and even surpasses the strong closed-source Gemini-3.0-flash baseline. Empirically, we observe that D-ORCA learns to segment spoken sentences at a finer temporal granularity within its generated captions, which directly contributes to its more accurate time localization. Additionally, the lower absolute IoU scores observed for Chinese videos reflect the inherent greater difficulty of segmenting fragmented Mandarin speech.

To further evaluate D-ORCA's dialogue-centric captioning capabilities, we conduct an additional assessment using a Caption-Based QA approach. We also test D-ORCA's performance across a wide range of general audio-visual understanding benchmarks. Detailed results and analyses are provided in Appendix D.

5.2 Impact of pre-DPO

Since the audio encoder is not native to the visual LLM backbone, the SFT model suffers from severe repetition degeneration when processing complex audio-visual inputs. As shown in Table 2, the SFT

Table 1: Results of different omni-modal LLMs on our curated DVD-Bench, including English (En) subset and Chinese (Zh) subset. The speaker attribution accuracy (Acc), global speech WER/CER, and sentence-level temporal IoU are evaluated. "-" denotes instances where valid metrics could not be computed (e.g., the model fails to generate timestamps or cannot perform Chinese ASR).

Model	DVD-Bench (En)			DVD-Bench (Zh)		
	Acc% ↑	WER% ↓	IoU% ↑	Acc% ↑	CER% ↓	IoU% ↑
WhisperX (Bain et al., 2023)	68.8	**15.2**	45.4	51.1	24.6	9.0
ARC-Qwen-Video-Narrator (7B) (Ge et al., 2025)	66.4	65.0	23.0	63.2	53.6	10.1
Qwen2.5-Omni (7B) (Xu et al., 2025a)	62.7	83.6	-	55.7	69.4	-
video-SALMONN 2+ (7B) (Tang et al., 2025)	66.6	94.0	-	59.9	-	-
AVoCaDO (7B) (Chen et al., 2025)	72.9	17.9	-	69.3	-	-
Qwen3-Omni-Instruct (30B-A3B) (Xu et al., 2025b)	67.8	91.3	-	63.5	60.6	-
Gemini-3.0-flash	**81.2**	15.8	39.5	**78.8**	**16.9**	24.1
Ours-SFT Model (8B)	71.2	29.8	31.8	69.6	30.3	24.7
D-ORCA (8B)	81.1	16.6	**57.1**	78.0	17.5	**37.8**

baseline exhibits a repetition rate of 21.8%. This pathology critically destabilizes GRPO: frequent length violations result in zero rewards for a certain portion of samples, causing gradient sparsity that hinders effective optimization.

To validate the necessity of pre-DPO, we conducted a comparative ablation on English scenarios. We compared applying GRPO directly to the SFT model (SFT → GRPO) versus our proposed pipeline (SFT → pre-DPO → GRPO). To isolate the impact, we excluded the temporal reward for this experiment ($\lambda_1 = 0.9, \lambda_2 = 0.1, \lambda_3 = 0$) and trained both settings for 200 steps. The results indicate that direct GRPO struggles to correct the decoding pathology, retaining a notable repetition rate of 4.3% and achieving suboptimal performance. In contrast, pre-DPO effectively regularizes the output distribution, reducing the repetition ratio to 0.7%. This allows the subsequent GRPO stage to focus entirely on fine-grained dialogue refinement, resulting in better speaker attribution and speech fidelity.

Table 2: Experiment on the impact of pre-DPO. Results are reported on DVD-Bench (En). "Rep%" denotes the repetition ratio of generated captions, and "Acc%" denotes speaker attribution accuracy.

Method	DVD-Bench (En)		
	Rep% ↓	Acc% ↑	WER% ↓
SFT	21.8	71.2	29.8
SFT → GRPO	4.3	78.9	20.4
SFT → pre-DPO	**0.7**	71.9	18.7
SFT → pre-DPO → GRPO	0.8	**79.4**	**18.5**

5.3 Ablation on Reward Balancing in GRPO

Optimizing for dialogue-centric video understanding requires a delicate balance between multiple objectives: identifying speakers, transcribing content, and localizing speech. We conducted ablation studies on English scenarios to determine the optimal configuration of reward coefficients. All experiments were conducted for 1,000 steps, about one epoch. Results are summarized in Table 3.

Table 3: Ablation study of reward balancing strategies. Evaluation is performed on DVD-Bench (En). "Collapse" indicates model degeneration where metrics cannot be computed.

Settings		DVD-Bench (En)		
		Acc% ↑	WER% ↓	IoU% ↑
pre-DPO Model		71.9	18.7	35.5
GRPO Variants ($\lambda_1 : \lambda_2 : \lambda_3$)				
1 : 0 : 0	w/o K_{warmup}	78.2	27.8	-
9 : 1 : 0	w/o K_{warmup}	80.5	16.5	-
9 : 1 : 1	w/o K_{warmup}	*Collapse at early stage (∼200 steps)*		
9 : 1 : 0.1	w/o K_{warmup}	*Collapse at mid stage (∼600 steps)*		
9 : 1 : 1	$K_{\text{warmup}} = 500$	79.8	16.8	55.1

Table 4: Robustness analysis of evaluation metrics using different external LLMs as parsers. The pre-DPO model and the final D-ORCA are evaluated on DVD-Bench (En).

Model	Parser LLM	DVD-Bench (En)		
		Acc% ↑	WER% ↓	IoU% ↑
pre-DPO	Gemini-2.5-flash	71.9	18.7	35.5
	Qwen3-235B-A22B	67.8	17.9	25.2
	GPT-4.1-mini	71.6	17.6	31.3
D-ORCA	Gemini-2.5-flash	81.1	16.6	57.1
	Qwen3-235B-A22B	78.0	16.6	47.9
	GPT-4.1-mini	79.9	16.2	41.4

Given that the pre-DPO model already possesses a reasonable ability to transcribe speech, the core challenge lies in correctly binding this speech to its speaker. Consequently, we identify the speaker attribution reward r_{speaker} as the primary driver of optimization. However, our experiments reveal that optimizing exclusively for speaker attribution (λ_1:λ_2:λ_3=1:0:0) over-optimizes this single objective, leading to speech content recognition deteriorating significantly. The WER increases by about 9% compared with the pre-DPO model. In addition, we also observe that the generation becomes unfluent and incoherent, suggesting the general linguistic capability of the model is impaired. Incorporating a small fraction of the content reward (λ_1:λ_2:λ_3=9:1:0) acts as a stable regularizer, maintaining a low WER of 16.5% while achieving a high speaker accuracy of 80.5%.

Incorporating temporal localization via r_{time} introduces additional optimization challenges. When the temporal reward is applied from the start of training, we observe training collapse at early ($\sim$200 steps) or intermediate ($\sim$600 steps) stages, even under balanced ($\lambda_1 : \lambda_2 : \lambda_3 = 9 : 1 : 1$) or reduced ($9 : 1 : 0.1$) weighting schemes. This instability likely stems from the fact that the temporal reward not only requires precise localization but also implicitly constrains the generation format. When the model is still learning the semantic associations of who and what is spoken, the additional when constraint substantially enlarges the optimization space, encouraging degenerate strategies that exploit anomalous output formats to "hack" the temporal reward, ultimately leading to training collapse.

To mitigate this, we employ a curriculum strategy and set $K_{\text{warmup}} = 500$, which is halfway through the epoch. This sequential approach introduces r_{time} only after the model's basic dialogue-centric learning stabilizes, and effectively prevents reward hacking and ensures convergence, boosting the temporal IoU to 55.1% without compromising speaker accuracy and speech recognition.

5.4 Robustness of LLM-based Evaluation Metrics

Since traditional n-gram metrics are inadequate for assessing the semantic granularity of detailed video captions, we leverage advanced external LLMs, such as Gemini-2.5-flash, to parse our dialogue-centric results for evaluation. Generally, utilizing LLMs as judges introduces potential risks, including stochasticity, sensitivity to prompt phrasing, and inherent biases toward certain output styles. However, in our framework, we mitigate these risks by decomposing the complex evaluation of dialogue-centric captioning into three atomic, objective extraction tasks: "when, who, and what is said." This decomposition transforms the evaluation from a subjective scoring task into a rigorous information extraction problem, which is well within the capabilities of modern LLMs.

To validate the reliability of our metrics, we compared the evaluation results on our English benchmark using three distinct powerful LLMs as parsers: Gemini-2.5-flash, Qwen3-235B-A22B, and GPT-4.1-mini. As shown in Table 4, the results for speaker attribution accuracy and speech WER are consistent across all LLMs, since extracting spoken content and identifying the corresponding speaker constitutes a fundamental reading comprehension task that is straightforward for advanced LLMs. Evaluating temporal IoU for each speech sentence introduces some variability. This likely stems from the fact that different LLMs exhibit distinct preferences when handling edge cases, such as partial timestamps (*e.g.*, a start time without an end time) or mismatches between the model's sentence segmentation and the ground truth. However, the relative performance trends remain identical. D-ORCA consistently outperforms the pre-DPO baseline by a significant margin regardless of the judge employed. This cross-model consistency confirms that our proposed metric is robust and provides a fair assessment of dialogue-centric audio-visual understanding. We show more details in Appendix E.

6 Conclusion

In this paper, we introduced D-ORCA, a dialogue-centric omni-modal LLM designed to resolve the "who, what, and when" of audio-visual narratives through a specialized reinforcement learning framework. To this end, we proposed three novel reward signals tailored for speaker attribution accuracy, speech recognition robustness, and temporal grounding, which simultaneously serve as effective evaluation metrics for dialogue-centric video captioning. By curating the high-quality dialogue-rich bilingual DVD-Train dataset and implementing a robust training pipeline, D-ORCA achieves precise grounding of speech to characters and timestamps. Experiments demonstrate that D-ORCA not only establishes superior performance on our curated DVD-Bench but also maintains competitive capabilities across general audio-visual understanding benchmarks. We believe that understanding character dialogue is fundamental to holistic audio-visual comprehension, and D-ORCA marks a significant step forward toward this goal.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets, implementing methods (specifically, writing scripts to make API requests), and assisting with translation. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, designing or providing feedback on research methodology or experiments, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code and editing the paper to improve readability. We have reviewed all AI-assisted work. Specifically, all code generated by large language models was thoroughly reviewed and verified through correctness testing by the authors. We take full responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep Audio-visual Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common Voice: A Massively-Multilingual Speech Corpus. In *Proc. LREC*, Marseille, 2020.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *Proc. Interspeech*, 2023.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark. In *Proc. ICLR*, Singapore, 2025.

Xinlong Chen, Yue Ding, Weihong Lin, Jingyun Hua, Linli Yao, Yang Shi, Bozhou Li, Yuanxing Zhang, Qiang Liu, Pengfei Wan, et al. AVoCaDO: An Audiovisual Video Captioner Driven by Temporal Orchestration. *arXiv preprint arXiv:2510.10395*, 2025.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-Holmes: Can MLLM Think Like Holmes for Complex Video Reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Ming Cheng and Ming Li. 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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Samuele Cornell, Matthew S Wiesner, Shinji Watanabe, Desh Raj, Xuankai Chang, Paola Garcia, Yoshiki Masuyam, Zhong-Qiu Wang, Stefano Squartini, and Sanjeev Khudanpur. The CHiME-7 DASR Challenge: Distant Meeting Transcription with Multiple Devices in Diverse Scenarios. In *Proc. CHiME*, Dublin, 2023.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An Audio Captioning Dataset. In *Proc. ICASSP*, Barcelona, 2020.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The First-ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. In *Proc. CVPR*, Nashville, 2025.

Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, et al. ARC-Hunyuan-Video-7B: Structured Video Comprehension of Real-world Shorts. *arXiv preprint arXiv:2507.20939*, 2025.

Mao-Kui He, Jun Du, and Chin-Hui Lee. End-to-end Audio-Visual Neural Speaker Diarization. In *Proc. Interspeech*, Incheon, 2022.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating Real-world Omnimodal Understanding for Multimodal LLMs. *arXiv preprint arXiv:2502.04326*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank Adaptation of Large Language Models. In *Proc. ICLR*, 2022.

Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, Tianyan Zhou, and Takuya Yoshioka. Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of any Number of Speakers. In *Proc. Interspeech*, Incheon, 2020.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating Captions for Audios in The Wild. In *Proc. NAACL-HLT*, Minneapolis, 2019.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Yixuan Li, Changli Tang, Jimin Zhuang, Yudong Yang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. Improving LLM Video Understanding with 16 Frames Per Second. In *Proc. ICML*, Vancouver, 2025.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, 2004.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. NVILA: Efficient Frontier Visual Language Models. *arXiv preprint arXiv:2412.04468*, 2024.

Liang Lu, Naoyuki Kanda, Jinyu Li, and Yifan Gong. Streaming Multi-Talker Speech Recognition with Joint Speaker Identification. In *Proc. Interspeech*, Brno, 2021.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-Captioner: Data Pipeline, Models, and Benchmark for Omni Detailed Perception. *arXiv preprint arXiv:2510.12720*, 2025.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. WavCaps: A ChatGPT-Assisted Weakly-labelled Audio Captioning Dataset for Audio-language Multimodal Research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024.

Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, et al. See, Hear, and Understand: Benchmarking Audiovisual Human Speech Understanding in Multimodal Large Language Models. *arXiv preprint arXiv:2512.02231*, 2025.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *Proc. ICASSP*, Brisbane, 2015.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proc. ACL*, Philadelphia, 2002.

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan. A Review of Speaker Diarization: Recent Advances with Deep Learning. *Computer Speech & Language*, Volume 72:101317, 2022.

Qiyue Qiu, Tong Xu, and Enhong Chen. Visual-Enhanced End-to-end Neural Diarization. In *Proc. IVSP*, 2022.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-scale Weak Supervision. In *Proc. ICML*, Honolulu, 2023.

Desh Raj, Pavel Denisov, Zhuo Chen, Hakan Erdogan, Zili Huang, Maokui He, Shinji Watanabe, Jun Du, Takuya Yoshioka, Yi Luo, et al. Integration of Speech Separation, Diarization, and Recognition for Multi-speaker Meetings: System Description, Comparison, and Analysis. In *Proc. SLT*, Shenzhen, 2021.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

Rahul Sharma and Shrikanth Narayanan. Using active speaker faces for diarization in tv shows. *arXiv preprint arXiv:2203.15961*, 2022.

Guangzhi Sun, Yudong Yang, Jimin Zhuang, Changli Tang, Yixuan Li, Wei Li, Zejun MA, and Chao Zhang. video-SALMONN-o1: Reasoning-enhanced Audio-visual Large Language Model. In *Proc. ICML*, Vancouver, 2025.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Extending Large Language Models for Speech and Audio Captioning. In *Proc. ICASSP*, Seoul, 2024.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Captioning-Enhanced Audio-Visual Large Language Models. *arXiv preprint arXiv:2506.15220*, 2025.

Quan Wang, Yiling Huang, Guanlong Zhao, Evan Clark, Wei Xia, and Hank Liao. DiarizationLM: Speaker Diarization Post-Processing with Large Language Models. In *Proc. Interspeech*, Kos Island, 2024.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VATEX: A Large-scale, High-quality Multilingual Dataset for Video-and-language Research. In *Proc. ICCV*, Seoul, 2019.

Peiran Wu, Yunze Liu, Zhengdong Zhu, Enmin Zhou, and Junxiao Shen. UGC-VideoCaptioner: An Omni UGC Video Detail Caption Model and New Benchmarks. *arXiv preprint arXiv:2507.11336*, 2025.

Abudukelimu Wuerkaixi, Kunda Yan, You Zhang, Zhiyao Duan, and Changshui Zhang. DyViSE: Dynamic vision-guided speaker embedding for audio-visual speaker diarization. In *Proc. MMSP*, Shanghai, 2022.

Eric Zhongcong Xu, Zeyang Song, Satoshi Tsutsui, Chao Feng, Mang Ye, and Mike Zheng Shou. AVA-AVD: Audio-Visual Speaker Diarization in the Wild. In *Proc. ACM MM*, Lisboa, 2022.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, et al. Qwen3-Omni Technical Report. *arXiv preprint arXiv:2509.17765*, 2025b.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Proc. CVPR*, Las Vegas, 2016.

Chenyu Yang, Mengxi Chen, Yanfeng Wang, and Yu Wang. Uncertainty-Guided End-to-End Audio-Visual Speaker Diarization for Far-Field Recordings. In *Proc. MM*, Ottawa, 2023.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric Video Understanding Benchmark Without Text Shortcut. In *Proc. EMNLP*, Suzhou, 2025.

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

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. OmniVinci: Enhancing Architecture and Data for Omni-Modal Understanding LLM. *arXiv preprint arXiv:2510.15870*, 2025.

Han Yin, Yafeng Chen, Chong Deng, Luyao Cheng, Hui Wang, Chao-Hong Tan, Qian Chen, Wen Wang, and Xiangang Li. SpeakerLM: End-to-end Versatile Speaker Diarization and Recognition with Multimodal Large Language Models. *arXiv preprint arXiv:2508.06372*, 2025.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. WenetSpeech: A 10000+ Hours Multi-domain Mandarin Corpus for Speech Recognition. In *Proc. ICASSP*, Singapore, 2022.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video Instruction Tuning with Synthetic Data. *arXiv preprint arXiv:2410.02713*, 2024.

Xianrui Zheng, Chao Zhang, and Phil Woodland. Tandem Multitask Training of Speaker Diarisation and Speech Recognition for Meeting Transcription. In *Proc. Interspeech*, Incheon, 2022.

Xianrui Zheng, Chao Zhang, and Phil Woodland. DNCASR: End-to-End Training for Speaker-Attributed ASR. In *Proc. ACL*, Barcelona, 2025.

Ziwei Zhou, Rui Wang, and Zuxuan Wu. Daily-Omni: Towards Audio-Visual Reasoning with Temporal Alignment across Modalities. *arXiv preprint arXiv:2505.17862*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, et al. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*, 2025.

A Details of Training Cost

The training settings and cost are shown in Table 5.

Table 5: Training settings and resource allocation for each stage of D-ORCA. #GPUs denotes the number of A800 GPUs used.

Stage	Batch Size	#Updates	#GPUs	Time Cost
Audio Alignment	1024	13,452	128	6h
Audio-Visual SFT	32	10,520	32	8h
pre-DPO	8	850	8	3h
GRPO	8	1,000	64	75h

B Details on Parsing Unilateral Unknown Timestamps

Extracting precise temporal intervals for spoken sentences from model-generated captions is not always straightforward due to the variability in generation styles. While an ideal output contains explicit start and end timestamps for every sentence, the model often produces ambiguous temporal

cues. To handle this, we define two specific scenarios where the external LLM parser is required to generate a "Unilateral Unknown" interval (i.e., $[\hat{t}_{start}, \text{Unk}]$ or $[\text{Unk}, \hat{t}_{end}]$).

Scenario 1: Discrete Time Points. The first scenario occurs when the model provides a single timestamp marking the onset of a sentence, rather than a full duration. For instance, the model might generate a caption such as: *"At 00:01, Judy says 'Sir, you were going 115 miles per hour.' "*. In this case, while the start time is explicitly defined, the end time is implicit. Consequently, the parser is configured to output the interval [00:01, Unk].

Scenario 2: Segmentation Granularity Mismatch. The second scenario arises when the segmentation of the generated text differs from the ground truth annotation, typically when the model aggregates multiple short sentences into a single block. Consider the following example:

- **Ground Truth:**
  - Sentence A (*"Sir, you were going 115 miles per hour."*): 00:01–00:03
  - Sentence B (*"I hope you have a good explanation."*): 00:04–00:05
- **Model Output:** *"00:01-00:05, Judy says 'Sir, you were going 115 miles per hour. I hope you have a good explanation.' "*

Here, the model provides a valid global time range (00:01 to 00:05) for the entire speech block, but fails to delineate the specific internal boundary where Sentence A ends and Sentence B starts. In such cases, we attribute the known boundaries to the respective sentences: the first sentence inherits the global start time, resulting in [00:01, Unk], while the last sentence inherits the global end time, resulting in [Unk, 00:05]. This approach allows us to partially credit the model for correctly identifying the boundaries of the dialogue exchange, even if the internal segmentation is coarse.

C Details of Our Curated DVD Datasets

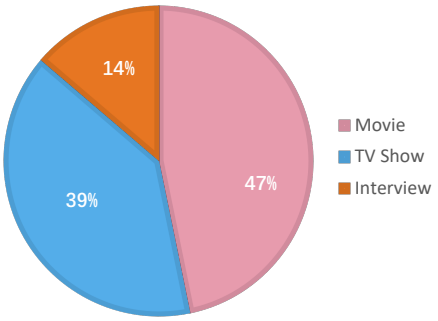
The proposed DVD dataset is constructed by collecting publicly available videos from platforms like YouTube and WeChat Channels. All selected videos were freely accessible to the public at the time of collection, and no paywalls, privacy settings, or security measures were bypassed. Notably, all data collected in this research are used strictly for non-commercial, academic research purposes under the Fair Use doctrine. We do not claim ownership of the video copyrights, which remain with their respective creators.

Due to the massive scale of the video collection, acquiring fine-grained human annotations for tens of thousands of multi-party dialogue videos is practically prohibitive. To facilitate scalable training, we construct the DVD-Train set by employing an automated annotation pipeline powered by a strong off-the-shelf multimodal model, e.g., Gemini-3.0-flash. It is important to note that our objective is not model distillation; rather, we utilize this pipeline strictly as a proxy annotator to generate large-scale pseudo-labels directly from the raw video files. These pseudo-labels are utilized in two distinct ways: providing initial descriptive video captions for the SFT warm-up stage, and extracting structured dialogue triplets (speakers, speech content, and timestamps) to formulate the reference targets for our GRPO training. This scalable paradigm provides the necessary data environment to validate our core algorithmic contribution: demonstrating that our proposed RL framework can systematically align cross-modal cues and substantially transcend the limitations of standard SFT.

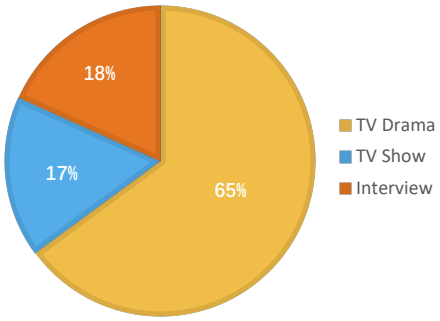
To ensure a comprehensive assessment of dialogue-centric audio-visual understanding, we compiled a robust evaluation benchmark, DVD-Bench, which consists of 964 English videos and 1,014 Chinese videos. These samples were meticulously curated to cover a wide spectrum of real-world conversational scenarios. The dataset spans from TV dramas, movies, TV shows, and interviews. Different from DVD-Train, DVD-Bench is refined and annotated by professional annotators, who are asked to give an accurate dialogue annotation of a video, e.g., write down the correct speaker of the speech content.

The specific distribution of video sources for each language is illustrated in Figure 3.

Detailed statistical specifications of the benchmark are provided in Table 6. This includes the total number of videos, the average duration, the average number of characters in the video, and the average number of spoken words.



(a) DVD-Bench (En)



(b) DVD-Bench (Zh)

Figure 3: Distribution of video sources in English (En) and Chinese (Zh) subset of our DVD-Bench. The dataset covers a diverse range of genres, including movies, TV dramas, TV shows, and interviews.

Table 6: Detailed statistics of the DVD-Bench.

Language	# Videos	Avg. Duration (s)	Avg. Characters	Avg. Words
English	964	83.9	3.5	156.9
Chinese	1014	94.6	5.0	330.9

D Additional Evaluation Results

Caption-Based QA Evaluation for Dialogue Understanding. While DVD-Bench has been carefully human-refined to guarantee evaluation quality, it is essential to demonstrate that the capabilities learned by D-ORCA generalize beyond our specific benchmark and that our proposed metrics are highly reliable. To evaluate performance from a different perspective, we employ a Caption-Based QA approach. Specifically, in this paradigm, each model first generates a detailed video caption. Then, a powerful external LLM (we use Gemini-3.0-flash) is tasked with answering multiple-choice questions based *solely* on the generated text captions, without access to the original video. Naturally, the more accurate and comprehensive the generated video captions are, the higher the accuracy of the Caption-Based QA will be.

Since our work primarily targets dialogue-centric scenes, we conducted experiments on the Speaker Recognition subset of AV-SpeakerBench (Nguyen et al., 2025). The evaluation results for different models are presented in Table 7, alongside their corresponding speaker attribution accuracy on DVD-Bench (En) for reference.

Table 7: Performance comparison on AV-SpeakerBench-Recognition using the Caption-Based QA approach, compared with speaker attribution accuracy on DVD-Bench (En).

Model	AV-SpeakerBench-Rec Caption QA (Acc%)	DVD-Bench (En) (Acc%)
Qwen3-Omni	43.8	67.8
AVoCaDO	56.4	72.9
Gemini-3.0-flash	63.3	81.2
Ours-SFT	51.4	71.2
D-ORCA	**62.0**	**81.1**

As shown in Table 7, the performance trends on this external QA task are highly consistent with the speaker attribution accuracy evaluated on DVD-Bench, which firmly validates the reliability and robustness of our proposed evaluation metrics. Furthermore, on this independent Caption-Based QA task, the final D-ORCA model achieves an improvement of over 10 percentage points compared to the SFT baseline, significantly outperforming existing open-source models. This provides strong evidence that the capabilities enhanced during our RL training phase genuinely generalize to external dialogue-centric scenarios, demonstrating the effectiveness of the proposed GRPO framework.

General Audio-Visual Understanding Performance. Moreover, D-ORCA exhibits robust generalization beyond dialogue-centric tasks. To verify that our specialized dialogue optimization does not compromise the model's foundational capabilities, we evaluated D-ORCA on a comprehensive suite of general video understanding benchmarks.

As shown in Table 8, D-ORCA only uses a maximum visual input of 128 frames (about 10,000 tokens) to deliver superior performance on comprehensive benchmarks such as Video-MME (Fu et al., 2025), WorldSense (Hong et al., 2025), AVUT (Yang et al., 2025), Video-Holmes (Cheng et al., 2025), DailyOmni (Zhou et al., 2025), and AV-SpeakerBench (Nguyen et al., 2025). D-ORCA consistently surpasses strong open-source baselines of similar size across all evaluated metrics. Impressively, it remains highly competitive even when compared with substantially larger frontier models like the 30B-parameter Qwen3-Omni-Instruct. These results confirm that integrating our low-token-rate audio pipeline with a strong visual LLM, followed by our RL training framework, yields a highly capable and versatile audio-visual foundation model.

Table 8: Results of different omni-modal LLMs on general audio-visual understanding benchmarks. For each model, we conduct evaluations following its official recommended settings.

Model	Video-MME	WorldSense	AVUT	Video-Holmes	DailyOmni	AV-SpeakerBench
OmniVinci (7B)	68.6	48.2	-	-	66.5	-
ARC-Qwen-Video-Narrator (7B)	62.4	45.1	-	43.2	63.2	40.2
Qwen2.5-Omni (7B)	64.3	47.8	66.3	43.7	62.7	42.3
video-SALMONN 2+ (7B)	**73.4**	50.9	69.5	46.9	71.8	51.6
AVoCaDO (7B)	65.9	49.9	70.0	47.2	69.8	51.6
Qwen3-Omni-Instruct (30B-A3B)	70.5	**54.0**	72.0	**54.1**	69.8	54.1
D-ORCA (8B)	72.9	53.7	**76.1**	48.5	**78.5**	**55.0**

E More Details of the Robustness of LLM-based Evaluation Metrics

To see the potential randomness of the metrics in LLM's parsing, we evaluated D-ORCA's speaker attribution accuracy on DVD-Bench (En) three times, using Gemini-2.5-Flash. The resulting scores were 81.05%, 81.06%, and 81.04%, showing minimal variation and confirming the stability of the LLM-evaluated metrics.

Besides, in Section 5.4, we demonstrated the robustness of different parser LLMs during evaluation. A natural follow-up question is whether using different parser LLMs to guide the GRPO training process yields similarly consistent results. The answer is affirmative.

To investigate this, we conducted an additional experiment following the setup in Section 5.2, but substituted the reward parser with the open-source Qwen3-235B-A22B model. We then performed a cross-evaluation against the Gemini-2.5-flash model on the DVD-Bench (En) dataset. The results, measuring speaker attribution accuracy (%Acc) and content word error rate (%WER), are summarized in Table 9.

Table 9: Cross-evaluation results on DVD-Bench (En). We compare models trained with different GRPO reward parsers (Gemini-2.5-flash vs. Qwen3-235B-A22B) and evaluate them using both parsers. Metrics reported are speaker attribution accuracy (%Acc) and content word error rate (%WER).

Model	Evaluated w/ Gemini-2.5-flash	Evaluated w/ Qwen3-235B-A22B
pre-DPO Model	71.9, 18.7	67.8, 17.9
GRPO w/ Gemini-2.5-flash	79.4, 18.5	77.6, 18.2
GRPO w/ Qwen3-235B-A22B	77.5, 15.3	77.6, 15.4

As shown in Table 9, utilizing the open-source Qwen3 model as the reward parser effectively optimizes the target model, achieving competitive performance. More importantly, the cross-evaluation reveals a highly consistent trend: models trained with rewards derived from Gemini-2.5-flash show significant performance gains when evaluated by Qwen3-235B-A22B, and vice versa. This strong cross-model

consistency indicates that extracting speaker-related information from textual captions is an objective task for advanced LLMs. It further validates the robustness and generalizability of our proposed metrics, suggesting that the improvements are not artifacts of a specific evaluator but reflect real enhancements in the model's capabilities.

F Ablation Study on Visual Modality

A potential concern in audio-visual speaker attribution is whether the model genuinely relies on cross-modal alignment (e.g., matching a voice to a speaking face or physical context) rather than defaulting to audio-only diarization heuristics. Because completely removing the visual input disrupts the fundamental instruction-following capabilities of multi-modal LLMs, varying the input frame rate serves as an effective method to ablate the density of visual information and verify its contribution.

We evaluated the speaker attribution accuracy on DVD-Bench (En) using 16, 32, and 128 input frames. For a typical 84-second video in our benchmark, 16 frames equate to sampling roughly one frame every 5 seconds, whereas 128 frames provide a much denser resolution of about one frame every 0.6 seconds. Because the API for Gemini does not support explicit control over frame extraction rates, we compared D-ORCA with the open-source Qwen3-Omni model. The results are presented in Table 10.

Table 10: Ablation study on the impact of visual frame counts. The metric reported is speaker attribution accuracy (Acc%) evaluated on DVD-Bench (En).

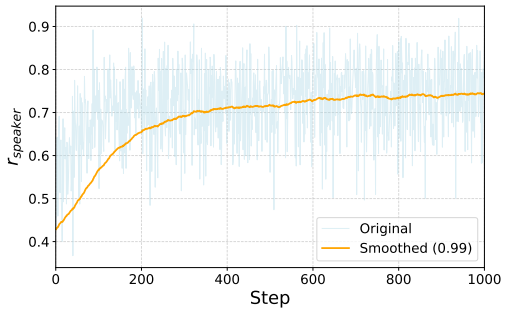
Model	**16 Frames**	**32 Frames**	**128 Frames**
Qwen3-Omni	61.7	64.0	65.1
D-ORCA	72.9	76.4	81.1

As shown in Table 10, reducing the number of input frames systematically decreases the speaker attribution capability for both models. This confirms that the visual modality is absolutely essential for this task. When visual sampling is too sparse, the model is likely to miss the brief moments when a specific character is speaking (e.g., due to camera cuts or off-screen context), making it impossible to correctly bind the voice track to the physical character.

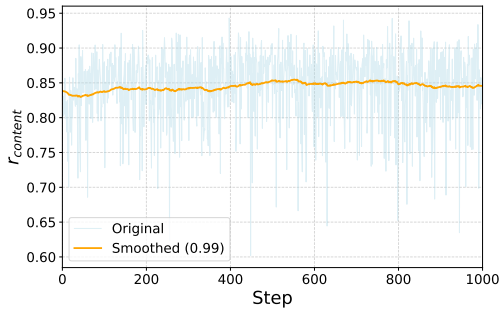
Furthermore, the models exhibit distinctly different scaling behaviors as visual information becomes denser. When the input frame count increases from 32 to 128, the performance improvement for Qwen3-Omni is relatively minor (+1.1%). In contrast, D-ORCA demonstrates a substantial and continuous performance gain (+4.7%) as visual density increases. This scaling trend strongly indicates that D-ORCA does not merely look at static frames, but successfully learns to leverage finer, high-resolution audio-visual alignment cues to assist in speaker judgment, firmly validating the cross-modal effectiveness of our proposed RL optimization framework.

G Training curves of GRPO

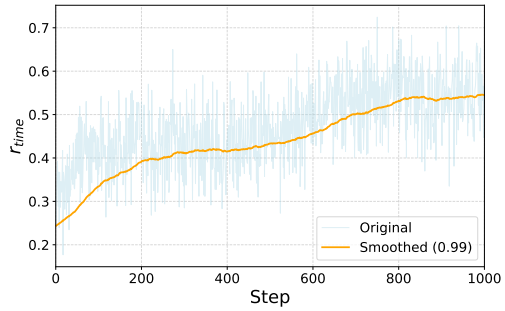
We visualize the training curves of the D-ORCA model during the GRPO stage for reproduction. Figure 4 illustrates the trajectories of the three individual reward components (r_{speaker}, r_{content}, r_{time}) as well as the total weighted reward (r_{total}).



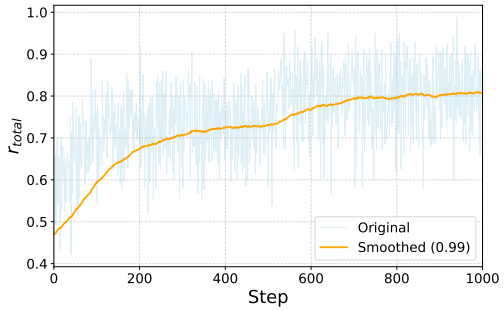
(a) Speaker Attribution Accuracy Reward (r_{speaker})



(b) Global Speech Content Reward (r_{content})



(c) Sentence-level Temporal Reward (r_{time})



(d) Total Weighted Reward (r_{total})

Figure 4: Training curves of D-ORCA during the GRPO stage. The plots demonstrate the progressive improvement of the model across (a) speaker attribution accuracy, (b) speech content fidelity, (c) sentence-level temporal precision, and (d) the overall objective reward function.