CastScript: Aligning "Who" and "What" in Multilingual Audio-Visual Captioning

Anonymous authors
Paper under double-blind review

Abstract

Speaker-attributed transcription, i.e., jointly determining who speaks and what is said, is a pivotal yet under-explored capability in audio-visual captioning, held back by two gaps. First, existing dialogue-centric benchmarks are largely limited to English and Chinese and lack scene and acoustic annotations for difficulty-stratified diagnosis. Second, speaker attribution and utterance transcription are heterogeneous competencies, yet existing methods optimize them under a single aggregate objective, yielding supervision that is neither dense nor dimension-specific. To close these gaps, we introduce **CastScriptBench** and **CastScript**. CastScriptBench comprises 508 in-the-wild conversational videos over 13 languages, spanning 1,356 shots, 4,236 utterances, and 1,218 speakers, annotated with 8 scene and acoustic attributes for difficulty-stratified analysis and evaluated by a verifiable protocol that combines edit-distance-based transcription evaluation with checklist-based speaker attribution. CastScript is the first to adapt multi-teacher on-policy distillation (MOPD) to captioning: unlike conventional MOPD that assigns experts by task, we route by output dimension within a single sequence, supervising tag tokens with an attribution expert and text tokens with a transcription expert, replacing the sequence-level scalar reward with per-token distribution matching. On CastScriptBench, both proprietary and open-source models mis-attribute and mistranscribe, worsening sharply as speaker count and overlap rise. CastScript surpasses even the transcription expert in fidelity, attains the best open-source performance, and establishes MOPD as an effective recipe for multi-dimensional captioning.

1 Introduction

Speaker-attributed transcription, jointly identifying *who* speaks (speaker attribution) and faithfully transcribing *what* is said (transcription fidelity), is a pivotal capability in audio-visual video captioning (Tang et al., 2025; Wu et al., 2025; Chen et al., 2026a). By tying each spoken utterance to its speaker, it produces richer, audio-grounded descriptions that measurably benefit downstream video understanding and generation (Chen et al., 2024a; Du et al., 2025; Chen et al., 2024b).

Unfortunately, progress on this capability is held back by two limitations. **1)** Existing benchmarks (Nguyen et al., 2026; Tang et al., 2026; Chen et al., 2026b) are coarse: they are largely limited to English and Chinese, limiting the coverage of linguistic and writing-system diversity in existing evaluations; and they characterise difficulty mainly by the number of speakers, offering no fine-grained analysis of how other acoustic and visual conditions, such as overlapping speech and background noise, or off-screen speakers and frequent shot changes, affect attribution and transcription. **2)** Existing methods post-train a pretrained multimodal model with supervised fine-tuning (SFT), reinforcement learning (RL), or both. SFT (e.g., Bai et al., 2026) is off-policy, imitating fixed references rather than the model's own errors, and its uniform token loss dilutes the few decision-critical tokens among many trivial ones. RL (e.g., Chen et al., 2026a;b; Tang et al., 2026) is on-policy but collapses each rollout into a single sequence-level scalar spread uniformly across tokens, so its signal is diluted and grounding errors become indistinguishable from transcription errors. In short, neither delivers the dense and dimension-specific supervision this task demands.

To address these limitations, we introduce **CastScriptBench**, a multilingual benchmark of 508 in-the-wild conversational videos over 13 languages (48.4% in non-Latin scripts), totalling 1,356 shots and 4,236 utterances from 1,218 speakers, each clip labelled with 8 scene and acoustic attributes for

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.

difficulty-stratified analysis. We deliberately retain clips where acoustic interference makes audio alone insufficient for attribution, so that the benchmark demands genuine audio-visual grounding. Each clip is paired with a structured two-part reference that supports verifiable per-dimension scoring: transcription by a character-level distance metric and attribution by a checklist against the reference. We further introduce **CastScript**: inspired by multi-teacher on-policy distillation (MOPD) for LLMs (Ma et al., 2026a), we adapt it to the captioning setting. Unlike prior MOPD that routes to teachers by *task* (e.g. math vs. coding), we exploit a property specific to captioning, namely that a single output sequence interleaves tokens reflecting different fine-grained perceptual dimensions, and route supervision by token type: tag tokens are supervised by an attribution expert and utterance tokens by a transcription expert. This makes the supervision at once *denser*, giving every token a full distributional target rather than one sequence-level scalar, and *dimension-specific*, matching each token to the expert responsible for its dimension; together they resolve the signal dilution that a single RL reward cannot. Built on Qwen2.5-Omni-7B (Xu et al., 2025a), CastScript unifies both experts within a single model, matching the attribution expert and even surpassing the transcription expert in fidelity; despite its modest size, it substantially narrows the gap to far larger proprietary systems. Our contributions are summarised as follows.

- We build **CastScriptBench**, a multilingual benchmark of 508 in-the-wild conversational videos over 13 languages whose attribution demands genuine audio-visual grounding, paired with structured references and 8 attributes for verifiable per-dimension and difficulty-stratified evaluation.
- We propose **CastScript**, trained via token-routed MOPD, the first multi-teacher on-policy distillation for captioning, which routes supervision by output dimension for dense, dimension-specific training.
- We benchmark and analyse the strongest proprietary and open-source multimodal models on CastScriptBench, on which our MOPD-trained CastScript attains stronger and better-balanced performance across the two dimensions than RL-based methods.

2 Related Work

2.1 Speaker-Attributed Transcription

Traditional speaker-attributed transcription typically combines automatic speech recognition with speaker diarisation (Park et al., 2022), or jointly models both components in an end-to-end system (Kanda et al., 2020; Yin et al., 2026). However, these approaches cannot directly associate speakers with visual descriptions. With the emergence of multimodal models, visual information can assist speaker attribution (Liu et al., 2022), while dedicated components and training designs further exploit cross-modal cues (Tang et al., 2026; Bai et al., 2026). These approaches nevertheless depend heavily on strong visual evidence and suffer from unstable cross-modal alignment, limiting robustness in multi-speaker videos with frequent shot changes. We instead explicitly train the audio-visual spatiotemporal alignment required for speaker attribution, enabling end-to-end speaker-attributed transcription in such scenarios.

2.2 Dialogue-Centric Audio-Visual Captioning Models

Omni-modal models such as Qwen-Omni (Xu et al., 2025a) and Gemini (Gemini Team, 2025), which can perceive both visual and auditory signals, have the ability to caption audio-visual content. Building on this, models such as UGC-VideoCaptioner (Wu et al., 2025), AVSCap (Wang et al., 2026), Omni-Captioner (Ma et al., 2026b), and AVoCaDO (Chen et al., 2026a) conduct post-training with dimension-specific data or rewards to further improve overall captioning quality. Human dialogue, which carries rich caption information, has since drawn increasing attention: D-ORCA (Tang et al., 2026) and DiaDem (Chen et al., 2026b) apply dialogue-centric, multi-dimensional rewards for further GRPO (Shao et al., 2024) training. However, these methods mix reward signals from different dimensions into a single composite sequence-level reward, leaving distinct output dimensions without targeted supervision and allowing signals from harder dimensions to be overwhelmed. Our token-level MOPD instead applies dimension-specific supervision to the corresponding tokens, directly addressing this limitation.

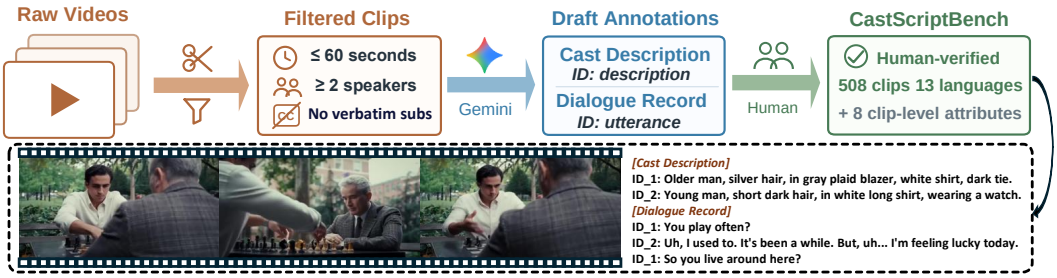


Figure 1: CastScriptBench construction pipeline and structured reference format.

3 CastScriptBench

CastScriptBench comprises 508 in-the-wild conversational clips spanning 13 languages and 7 writing systems (48.4% non-Latin), with 1,356 shots and 4,236 utterances from 1,218 speakers recorded under diverse real-world acoustic and visual conditions. Each clip is paired with a structured two-part reference: a *cast description* that binds every speaker label to a visual description, and a *dialogue record* of label-prefixed utterances in temporal order. This format lets us score speaker attribution and transcription independently, while eight scene and acoustic attributes per clip turn difficulty into an explicit, queryable property that supports fine-grained, difficulty-stratified diagnosis.

3.1 Data Sources

We build on two publicly released corpora that already contain unscripted multi-party speech with visible speakers. MSDWILD (Liu et al., 2022) contributes short, socially dense exchanges recorded in the wild, where speakers interrupt one another and the camera does not reliably follow whoever is talking. AVA (Gu et al., 2018), drawn from feature films across many production languages, contributes long-form cinematic dialogue with deliberate shot editing and non-English speech.

3.2 Construction Pipeline

Segmentation and filtering. Source videos are cut into clips no longer than 60 seconds. Using Qwen3-Omni-Instruct (Xu et al., 2025b), we discard clips with fewer than two speakers and those with same-language verbatim burned-in subtitles.

Reference annotation. Gemini 3.5 Flash (Google DeepMind, 2026) first produces a structured draft, which is corrected by one annotator, independently verified by a second, and adjudicated by a senior annotator when necessary. Eight clip-level attributes are independently labelled by three annotators and aggregated by majority vote. Further details are provided in Appendix B.1.

Checklist construction. From the human-corrected annotations shown in Figure 1, we derive one speaker–utterance verification item from each reference utterance. Repeated utterances are disambiguated using adjacent-utterance context. These checklist items serve as the basis for our subsequent evaluation. Further details are provided in Appendix B.2.

3.3 Data Analysis

Data composition and real-world challenges. To evaluate speaker attribution under realistic and diverse conditions (Figure 2), CastScriptBench spans 13 languages across seven writing systems; critically, nearly half of the clips (48.4%) use non-Latin scripts, exposing the turn-segmentation failures that Latin-centric benchmarks leave unmeasured. Visually, we cross indoor and outdoor settings with multi-shot editing (38.4% of clips), isolating the case in which a speaker leaves the frame from the fixed-camera case. Acoustically, 70.1% of clips carry at least one adverse factor—overlapping speech, background noise or background music. The benchmark contains 1,356 shots and 4,236 annotated turns from 1,218 distinct speakers, 97.6% of whom appear on screen, which keeps genuinely off-screen speech distinguishable from hallucinated attribution.

Multi-dimensional difficulty stratification. Rather than relying on a single aggregate metric that obscures specific failure modes, we count seven adverse factors for each clip—three or more speakers,

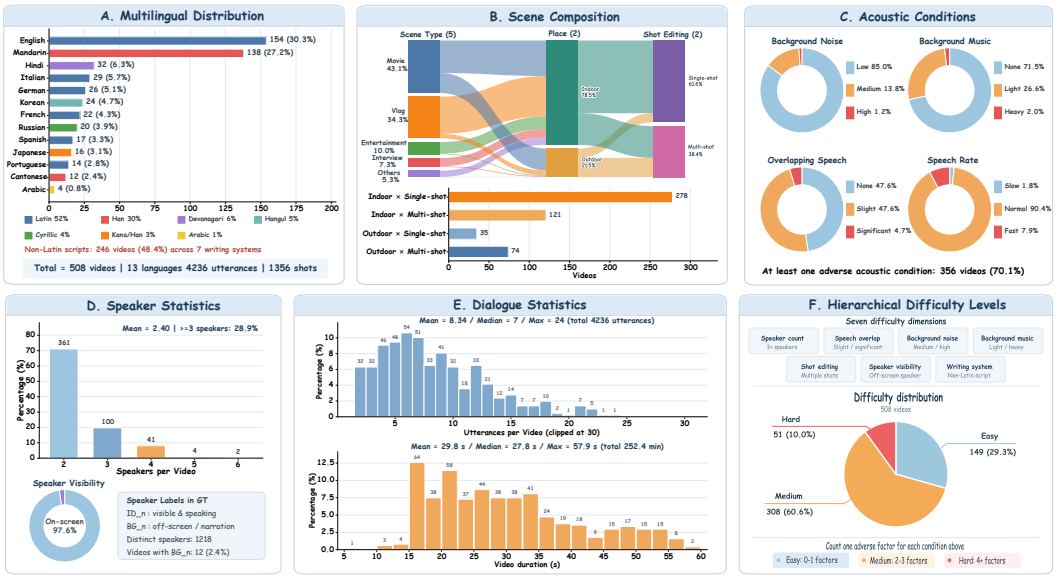


Figure 2: Composition of **CastScriptBench**. (A) Language distribution over 13 languages and 7 writing systems. (B) Scene type, setting and shot editing, with their cross-tabulation. (C) Acoustic conditions. (D) Speaker counts and visibility. (E) Utterances per clip and clip duration. (F) Hierarchical difficulty levels induced by seven adverse factors.

speech overlap, background noise, background music, multi-shot editing, an off-screen speaker, and a non-Latin writing system—and group clips into three macro difficulty tiers by the number of co-occurring factors: EASY (zero or one factor), MEDIUM (two or three), and HARD (four or more). These tiers serve as the primary testbed for overall model comparison (Table 1).

3.4 EVALUATION PROTOCOL

Because the accuracy of a dialogue description has two components—correct speaker attribution and precise utterance transcription—a prediction is evaluated separately along these two coupled axes.

Attribution accuracy. To evaluate speaker attribution, we feed the checklist obtained from the pipeline in Section 3.2, together with the model's generated response, into an LLM judge, so that a single call scores an entire clip; we use Qwen3.8-27B (Qwen Team, 2026b) as the judge. Let V be the set of clips, where clip v contains n_v dialogue turns and hence n_v checklist items. The judge scores question i of clip v against the prediction as

$$s_i^v = \begin{cases} 1 & \text{same-language content match and correct speaker attribution,} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

Specifically, a question scores 1 when the response contains a same-language, semantically equivalent utterance spoken by the person matching the correct cast description, and 0 otherwise—for instance, an incorrect speaker attribution or a missing utterance. Accuracy is averaged level by level, first within a clip and then over clips:

$$\mathrm{Acc}(v) = \frac{1}{n_v} \sum_{i=1}^{n_v} s_i^v, \qquad \mathrm{Acc} = \frac{1}{|V|} \sum_{v \in V} \mathrm{Acc}(v). \tag{2}$$

Transcription fidelity. To strictly evaluate verbatim transcription beyond attribution, we compute character-level Levenshtein similarity (Levenshtein, 1966). Let $\tau(\cdot)$ strip speaker labels, concatenate utterances in order, and apply standard text normalisation (NFKC, lowercasing, and stripping whitespace/punctuation). With $t_v = \tau(y_v)$ and $\hat{t}_v = \tau(\hat{y}_v)$:

$$\mathrm{Fid}(v) = \max\left(0,\ 1 - \frac{\mathrm{Lev}(\hat{t}_v,\ t_v)}{|t_v|}\right), \quad \mathrm{Fid} = \frac{1}{|V|} \sum_{v \in V} \mathrm{Fid}(v). \tag{3}$$

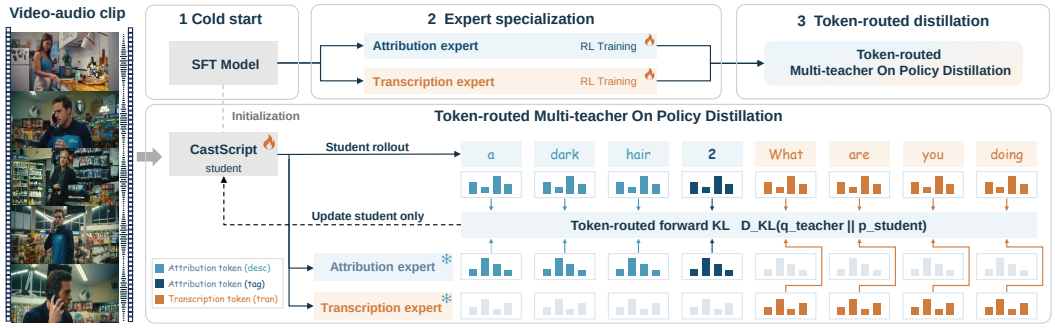


Figure 3: Training pipeline of **CastScript** and illustration of token-level MOPD.

4 CastScript

As illustrated in Figure 3, CastScript is built on Qwen2.5-Omni-7B (Xu et al., 2025a) and trained in three stages: (1) a supervised cold-start that formats the cast-then-dialogue structure, (2) single-objective reinforcement learning to obtain two specialized experts, and (3) multi-teacher on-policy distillation with within-sequence token-type routing.

4.1 Stage 1: Supervised Cold-Start

To establish the structured output syntax and initial audio-visual alignment, we fine-tune the base model using standard cross-entropy loss:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(X,Y)\sim\mathcal{D}_{\text{SFT}}}\left[\sum_{t=1}^{T}\log \pi_\theta(y_t \mid y_{<t}, X)\right]. \tag{4}$$

The resulting checkpoint π_{θ_0} outputs parseable cast descriptions and dialogue turns, serving as the common initialization for subsequent stages.

4.2 Stage 2: Single-Objective Expert Specialization

Speaker-attributed transcription demands two distinct fine-grained competencies: spatio-temporal audio-visual grounding for speaker identification, and multilingual acoustic recognition for verbatim transcription. We train two specialized experts independently from π_{θ_0} using GRPO (Shao et al., 2024):

$$\pi_{\theta_{\text{attr}}} = \arg\max_{\pi}\ \mathbb{E}_{X,\,\hat{y}\sim\pi}[R_{\text{attr}}(\hat{y}, y)]\,, \qquad \pi_{\theta_{\text{trans}}} = \arg\max_{\pi}\ \mathbb{E}_{X,\,\hat{y}\sim\pi}[R_{\text{trans}}(\hat{y}, y)]\,, \tag{5}$$

where R_{attr} optimizes attribution accuracy (Eq. 2) and R_{trans} optimizes transcription fidelity (Eq. 3). The experts provide dimension-specific supervision and reference performance for the *who* and *what* dimensions, respectively.

4.3 Stage 3: Multi-Teacher On-Policy Distillation with Token Routing

Why not mix-RL? The standard way to fold several capability dimensions into one model is to optimise a single weighted reward over all of them (Tang et al., 2026; Chen et al., 2026b),

$$R_{\text{mix}}(\hat{y}, y) \;=\; \sum_{k=1}^{K} w_k\, R_k(\hat{y}, y), \qquad \sum_{k=1}^{K} w_k = 1, \tag{6}$$

which for speaker-attributed transcription instantiates R_{attr} and R_{trans} as two equally weighted components; we refer to this as *mix-RL* and use it as our main training baseline.

However, speaker tags and utterance text are interleaved token by token in the output yet demand fundamentally distinct competencies: associating a tag is a cross-modal grounding problem (binding visual appearance and voice identity), whereas transcribing an utterance requires multilingual acoustic recognition. In GRPO (Shao et al., 2024), the reward of a rollout is collapsed into a single

group-normalised advantage scalar that multiplies the log-probability of every sampled token:

$$\nabla_\theta \mathcal{J}_{\text{mix}}(\theta) \;=\; \mathbb{E}_{X,\,\hat{y}\sim\pi_\theta}\left[A(\hat{y})\sum_{t=1}^{|\hat{y}|}\nabla_\theta \log \pi_\theta(\hat{y}_t \mid \hat{y}_{<t}, X)\right], \qquad A(\hat{y}) = \frac{R_{\text{mix}}(\hat{y}, y) - \mu}{\sigma}, \quad (7)$$

where μ and σ are the mean and standard deviation of rewards in the rollout group. The same advantage coefficient is assigned to tag and text tokens, although their token-specific gradients differ. The scalar does not explicitly identify whether a deficient output needs a corrected speaker label or a corrected transcript. In our experiments (Section 5), mix-RL underperforms each specialist on its target metric. We therefore replace this shared reward-derived coefficient with expert target distributions selected according to each token's role.

Token-routed distillation. To unify both expert capabilities into a single deployable model, we employ Multi-Teacher On-Policy Distillation (MOPD) (Ma et al., 2026a). As each generated sample follows a structured cast-then-dialogue format, we partition the token sequence into three disjoint sets: $\mathcal{I}_{\text{desc}}$ (cast descriptions and structural tokens), $\mathcal{I}_{\text{tag}}$ (speaker ID digits in dialogue turn prefixes), and $\mathcal{I}_{\text{tran}}$ (dialogue utterance transcripts).

During distillation, the student π_θ generates rollouts $\hat{y} \sim \pi_\theta(\cdot \mid X)$, and a token-type router assigns per-token supervision to the corresponding specialized expert:

$$\pi^{(t)}_{\text{teacher}}(\cdot) = \begin{cases} \pi_{\theta_{\text{attr}}}(\cdot \mid \hat{y}_{<t}, X) & \text{if } t \in \mathcal{I}_{\text{desc}} \cup \mathcal{I}_{\text{tag}}, \\ \pi_{\theta_{\text{trans}}}(\cdot \mid \hat{y}_{<t}, X) & \text{if } t \in \mathcal{I}_{\text{tran}}, \end{cases} \quad (8)$$

where both teachers share the student prefix $\hat{y}_{<t}$. The student minimizes token-wise forward KL between the selected teacher distribution q_t and the student distribution $p_{\theta,t}$, using a top-k support approximation detailed in Appendix C.1:

$$\mathcal{L}_{\text{MOPD}}(\theta) = \mathbb{E}_{X,\hat{y}\sim\pi_\theta}\left[\frac{1}{Z}\sum_{t=1}^{|\hat{y}|} w_t D_{\text{KL}}(q_t \,\|\, p_{\theta,t})\right], \quad (9)$$

where $w_t = w_{\text{tag}}$ for critical speaker ID tokens ($t \in \mathcal{I}_{\text{tag}}$), and $w_t = 1$ for all other tokens ($t \in \mathcal{I}_{\text{desc}} \cup \mathcal{I}_{\text{tran}}$), with normalization factor $Z = \sum_t w_t$. The sampled rollouts are held fixed during each distillation update.

Token routing assigns each position an explicit target from the expert trained for its output dimension. In our experiments, CastScript matches the attribution expert on Acc while surpassing the transcription expert on Fid, combining their complementary strengths in a single model.

5 Experiments

Section 5.1 benchmarks CastScript against representative proprietary and open-source models on CastScriptBench, followed by stratified analyses across benchmark difficulty factors. Section 5.2 ablates the proposed training pipeline and token-routing design, while Section 5.3 evaluates the reliability of our evaluation protocol and investigates key distillation choices. Complementary details are provided in the appendices: Appendix B documents benchmark construction; Appendix C reports training hyperparameters, training-data construction, and complete tables and analyses for diagnostic experiments and the external AV-SpeakerBench evaluation; Appendix D presents qualitative examples; and Appendix E provides all prompts used in this work. Appendix A discusses limitations.

5.1 Main Results and Stratified Analysis on CastScriptBench

We design extensive experiments to evaluate the effectiveness of our method. Table 1 measures performance across CastScriptBench difficulty tiers for three proprietary models: Qwen3.5-Omni-Plus (Qwen Team, 2026a), Gemini 3.5 Flash (Google DeepMind, 2026), and Doubao-Seed-2.0-lite (ByteDance Seed, 2026); eight open-source models: UGC-VideoCaptioner (Wu et al., 2025), Qwen2.5-Omni (Xu et al., 2025a), ASID-Captioner (Li et al., 2026), HumanOmniV2 (Yang et al., 2025), video-SALMONN 2+ (Tang et al., 2025), AVoCaDO (Chen et al., 2026a), D-ORCA (Tang et al., 2026), and Qwen3-Omni (captioner) (Xu et al., 2025b); and our CastScript. For deeper analysis, we select three representative models: Gemini 3.5 Flash, Qwen3-Omni (captioner), and CastScript. Table 2 examines speaker count and speech overlap, while Table 3 provides fine-grained results across acoustic, visual, and linguistic conditions.

Table 1: Speaker-attributed transcription on **CastScriptBench**. Acc: attribution accuracy (Eq. 2); Fid: transcription fidelity (Eq. 3).

Model	Size	Accuracy (Acc) ↑				Fidelity (Fid) ↑			
		Overall	Easy	Med	Hard	Overall	Easy	Med	Hard
Qwen3.5-Omni-Plus	—	0.798	0.843	0.783	0.759	0.911	0.944	0.904	0.863
Gemini 3.5 Flash	—	0.793	0.859	0.785	0.646	0.852	0.930	0.838	0.706
Doubao-Seed-2.0-lite	—	0.761	0.827	0.747	0.648	0.834	0.912	0.813	0.737
UGC-VideoCaptioner	3B	0.107	0.094	0.128	0.022	0.209	0.202	0.234	0.081
Qwen2.5-Omni	7B	0.080	0.045	0.107	0.019	0.182	0.137	0.220	0.086
ASID-7B	7B	0.068	0.128	0.046	0.031	0.139	0.256	0.095	0.056
HumanOmniV2	7B	0.016	0.025	0.013	0.007	0.076	0.087	0.073	0.064
video-SALMONN 2+	7B	0.002	0.001	0.003	0.000	0.012	0.008	0.014	0.007
AVoCaDO	7B	0.411	0.565	0.347	0.345	0.593	0.805	0.511	0.466
D-ORCA	8B	0.532	0.613	0.517	0.383	0.584	0.694	0.559	0.418
Qwen3-Omni	30B-A3B	0.586	0.619	0.597	0.421	0.676	0.765	0.665	0.485
CastScript (Ours)	**7B**	**0.650**	**0.730**	**0.641**	**0.467**	**0.814**	**0.893**	**0.798**	**0.683**

Performance comparison. As shown in Table 1, CastScript achieves the best overall performance among open-source models. Compared with the 30B-A3B Qwen3-Omni (captioner), CastScript (7B) improves Acc by 0.064 and Fid by 0.138. It also narrows the gap to proprietary models: its Fid of 0.814 is only 0.038 below Gemini 3.5 Flash and 0.020 below Doubao-Seed-2.0-lite. Across difficulty levels, all models with non-trivial performance exhibit a consistent monotonic decline from Easy to Hard, validating the effectiveness and discriminative power of the CastScriptBench difficulty partition. For most models, the relative degradation in Acc is greater than that in Fid. Compared with transcription, which primarily relies on unimodal audio decoding, speaker attribution requires more challenging cross-modal spatiotemporal alignment in complex multi-speaker scenarios.

Effect of Speaker Count and Speech Overlap. In complex multi-speaker scenarios, speaker count and speech overlap are the two primary interference factors that undermine cross-modal alignment. As both factors increase, all three models exhibit pronounced joint degradation (Table 2). Qwen3-Omni (captioner) drops to only 0.194 Acc in the most challenging setting. Under ≥ 4 speakers with significant overlap, however, CastScript retains 0.422 Acc and 0.602 Fid, surpassing Gemini 3.5 Flash (0.402 Acc / 0.554 Fid). These results demonstrate that an effective training recipe can equip an open-source model with stronger structural alignment than a proprietary model under extreme multi-speaker interference. The left example in Figure 4 shows a representative multi-speaker misalignment shared by the two comparison models.

Effect of Audio-Visual and Linguistic Conditions. Beyond speaker count and overlap, Table 3 reveals distinct model behaviours across diverse factors. Gemini 3.5 achieves the highest overall performance ceiling, showing exceptional strength under standard, clean conditions such as Latin-script clips and low noise. Qwen3-Omni (captioner) performs strongly in specific niches, such as slow speech rates. Crucially, CastScript outperforms Qwen3-Omni (captioner) across most conditions and shows stronger resilience than Gemini 3.5 under several challenging conditions. When visual cues are unavailable for off-screen speech, CastScript's Acc drops by only 9.8 percentage points, whereas Gemini 3.5 drops by 17.1 points. Under light background music, CastScript's Acc decreases by only 1.6 points, compared with 5.0 points for Gemini 3.5. Overall, while acoustic interference generally has a greater impact than visual factors, CastScript exhibits the strongest robustness in maintaining high performance under adverse real-world conditions. Nevertheless, the middle and right examples in Figure 4 illustrate that off-screen speech and fast speech can still cause attribution and transcription errors.

Overall. CastScript leads open-source models overall and remains particularly robust when speaker count and overlap increase. The results identify cross-modal speaker alignment, rather than transcription alone, as the central challenge of CastScriptBench. To further validate our method, we evaluate different models on AV-SpeakerBench by first generating structured captions and then using an LLM to answer questions from those captions. CastScript remains the strongest open-source model; see Appendix C.6.



Figure 4: Common failure cases in speaker-attributed transcription: multi-speaker misalignment, off-screen speaker misattribution, and fast-speech transcription errors (left to right). Green marks correct annotations and red marks errors.

Table 2: Attribution accuracy / transcription fidelity on the speaker-count × overlap grid.

Speakers	Model	Speech Overlap (Acc ↑ / Fid ↑)		
		None	Slight	Signif.
2	Gemini 3.5 Flash	0.832 / 0.902	0.823 / 0.836	0.752 / 0.827
	Qwen3-Omni	0.637 / 0.704	0.634 / 0.701	0.516 / 0.476
	CastScript (Ours)	**0.710 / 0.844**	**0.691 / 0.810**	**0.645 / 0.739**
3	Gemini 3.5 Flash	0.758 / 0.900	0.718 / 0.764	0.569 / 0.603
	Qwen3-Omni	0.513 / 0.613	0.533 / 0.725	0.361 / 0.512
	CastScript (Ours)	**0.561 / 0.812**	**0.591 / 0.807**	**0.425 / 0.605**
≥ 4	Gemini 3.5 Flash	0.750 / 0.926	0.753 / 0.829	0.402 / 0.554
	Qwen3-Omni	0.519 / 0.720	0.355 / 0.491	0.194 / 0.301
	CastScript (Ours)	**0.528 / 0.891**	**0.396 / 0.755**	**0.422 / 0.602**

Table 3: Attribution accuracy (Acc) and transcription fidelity (Fid) stratified by seven factors beyond speaker count and overlap. n denotes the number of clips per category.

			Gemini 3.5 Flash		Qwen3-Omni		CastScript	
Attribute	Category	n	Acc	Fid	Acc	Fid	Acc	Fid
Writing system	Latin	262	0.836	0.925	0.614	0.747	**0.676**	**0.875**
	Non-Latin	246	0.747	0.773	0.555	0.602	**0.622**	**0.750**
Shot editing	Single	313	0.793	0.836	0.575	0.700	**0.653**	**0.813**
	Multi	195	0.793	0.877	0.602	0.639	**0.644**	**0.817**
Setting	Indoor	399	0.797	0.842	0.583	0.692	**0.657**	**0.814**
	Outdoor	109	0.778	0.885	0.596	0.618	**0.622**	**0.818**
Noise level	Low	432	0.807	0.866	0.596	0.688	**0.666**	**0.827**
	Medium	70	0.715	0.768	0.537	0.628	**0.573**	**0.748**
	High	6	0.662	0.776	**0.444**	0.408	0.404	**0.699**
Background music	None	363	0.812	0.864	0.586	0.700	**0.659**	**0.830**
	Light	135	0.762	0.824	0.594	0.620	**0.643**	**0.783**
	Heavy	10	0.506	0.788	**0.482**	0.588	0.403	**0.670**
Speech rate	Slow	9	0.561	0.761	**0.720**	**0.650**	0.500	0.624
	Normal	459	0.806	0.861	0.597	0.688	**0.661**	**0.826**
	Fast	40	0.689	0.758	0.426	0.555	**0.551**	**0.728**
Speaker visibility	On-screen	496	0.797	0.852	0.589	0.675	**0.652**	**0.816**
	Off-screen	12	0.626	0.811	0.469	0.721	**0.554**	**0.760**

5.2 Ablation Study

We first perform supervised cold-start training on 18k examples, enabling the base model to acquire the fundamental capability required for this task. We then apply GRPO to the SFT model on 1k medium-difficulty examples using different reward configurations; training-data details are provided in Appendix C.2.

As shown in Table 4, jointly optimizing the two rewards in Mix-RL fails to match either single-objective expert on its target dimension. This confirms our analysis that a mixed scalar reward cannot

Table 4: Ablation on post-training strategies initialized from CastScript-SFT (π_{θ_0}).

Configuration	Training Strategy	Acc ↑	Fid ↑
Base (Qwen2.5-Omni-7B)	—	0.080	0.182
CastScript-SFT	Supervised fine-tuning	0.604	0.744
+ Continued SFT	SFT on the 1k RL-stage samples	0.623	0.748
Attribution Expert	RL with R_{attr}	**0.651**	0.768
Transcription Expert	RL with R_{trans}	0.583	<u>0.808</u>
Mix-RL	RL with $R_{\text{attr}} + R_{\text{trans}}$	0.620	0.791
MOPD w/o Routing	Uniform teacher assignment	0.637	0.797
CastScript (Ours)	**Token-routed teacher assignment**	<u>0.650</u>	**0.814**

provide targeted supervision for distinct output dimensions and yields smaller gains on the harder dimension that requires audio-visual spatiotemporal alignment. We therefore consolidate the two experts into the SFT model through MOPD and route their supervision to the token types associated with their respective dimensions. Token-routed MOPD nearly matches the attribution expert in Acc (0.650 vs. 0.651) and surpasses the transcription expert in Fid (0.814 vs. 0.808); more accurate preceding speaker tags may give the transcription expert clearer context for its per-token supervision. For the no-routing control, we keep the same training hyperparameters but average the two teachers' KL losses at every token. The results show that the control yields lower joint performance (0.637 Acc / 0.797 Fid).

To validate our post-training recipe, we further fine-tune the SFT model on the same 1k examples, but it remains substantially behind CastScript (0.623 Acc / 0.748 Fid vs. 0.650 Acc / 0.814 Fid). These results show that, in audio-visual captioning, targeted supervision of tokens from different dimensions improves model capability more effectively than a single mixed scalar objective.

5.3 Diagnostic Analysis

Evaluation Reliability. To evaluate the robustness of our evaluation system and its consistency with humans, we use the stronger Kimi-K3 (Kimi Team, 2026) and Qwen3.8-Max (Qwen Team, 2026b) to rescore the same model outputs. We also sample 100 clips and their checklists in proportion to the Easy, Medium, and Hard subsets for human evaluation. The three LLM judges produce highly correlated model rankings (pairwise Spearman $\rho \geq 0.993$; Spearman, 1904). Their majority vote reaches a mean Cohen's κ (Cohen, 1960) of 0.874 and 94.5% agreement with human annotations. Per-judge scores and further details are provided in Appendix C.3.

Divergence Objective and Tag Weighting. We examine the effect of different divergence objectives on training. When the teacher has moderate capability and produces uncertain distributions, forward KL outperforms reverse KL and JSD. For the more challenging speaker attribution dimension, which requires audio-visual alignment, appropriately increasing its supervision signal effectively improves performance. Complete tables and detailed analyses are provided in Appendices C.4 and C.5.

6 Conclusion

We introduced CastScriptBench, a speaker-attributed transcription benchmark comprising 508 in-the-wild conversational videos over 13 languages, spanning 1,356 shots, 4,236 utterances, and 1,218 speakers. It provides 8 scene and acoustic attributes, addressing the limited linguistic and attribute coverage of existing benchmarks. We further introduced CastScript, the first captioning model to use token-routed MOPD, which separates supervision for speaker grounding and utterance transcription. Experiments show that CastScript achieves the best open-source performance and stronger robustness than proprietary models under challenging conditions. Overall, we introduce a comprehensive benchmark tailored to speaker-attributed transcription, enabling fine-grained diagnosis across languages and complex real-world scenarios. We also validate the applicability of token-level MOPD to multi-dimensional captioning.

AI use statement

In this work, we used generative AI tools for drafting training and benchmark annotations, and for cleaning and reformatting datasets. We have not used generative AI tools for proposing or refining hypotheses, designing research methodology or experiments, or interpreting results. Additionally, we used generative AI tools for creating and editing software code and editing the paper to improve readability. All AI-generated code was manually reviewed and verified for correctness by the authors. All AI-drafted text was carefully checked, revised, and approved by the authors before inclusion. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

CastScriptBench builds on MSDWILD (Liu et al., 2022) and AVA (Gu et al., 2018), whereas our training clips are collected from publicly accessible online platforms. We intend to make the benchmark available exclusively for academic research under a restrictive license. Its release will take into account privacy concerns related to faces, voices, and dialogue and remain subject to the source materials' terms and applicable copyright requirements.

Reproducibility statement

We describe the benchmark sources, construction pipeline, and evaluation metrics in Section 3, with annotation and checklist details in Appendix B. Section 4 and Appendix C document the training procedure and experimental settings. Prompts are provided in Appendix E. Appendix C.3 additionally reports agreement across alternative judges and human annotations. Our planned public release will include CastScriptBench and our implementation. We are preparing the accompanying documentation to help others reproduce our experiments and build on these resources.

References

Detao Bai, Xihan Wei, and Zhiheng Ma. HumanOmni-Speaker: Identifying who said what and when. In *European Conference on Computer Vision (ECCV)*, 2026.

ByteDance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. ShareGPT4Video: Improving video understanding and generation with better captions. In *Advances in Neural Information Processing Systems*, volume 37, pp. 19472–19495, 2024a. doi: 10.52202/079017-0614.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70M: Captioning 70M videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13320–13331, 2024b.

Xinlong Chen, Yue Ding, Weihong Lin, Jingyun Hua, Linli Yao, Yang Shi, Bozhou Li, Qiang Liu, Yuanxing Zhang, Pengfei Wan, and Liang Wang. AVoCaDO: An audiovisual video captioner driven by temporal orchestration. In *International Conference on Learning Representations (ICLR)*, 2026a.

Xinlong Chen, Weihong Lin, Jingyun Hua, Linli Yao, Yue Ding, Bozhou Li, Bohan Zeng, Yang Shi, Qiang Liu, Yuanxing Zhang, Pengfei Wan, and Liang Wang. DiaDem: Advancing dialogue descriptions in audiovisual video captioning for multimodal large language models. In *European Conference on Computer Vision (ECCV)*, 2026b.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Yang Du, Zhuoran Lin, Kaiqiang Song, Biao Wang, Zhicheng Zheng, Tiezheng Ge, Bo Zheng, and Qin Jin. VC4VG: Optimizing video captions for text-to-video generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1124–1138, 2025. doi: 10.18653/v1/2025.emnlp-main.59.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Google DeepMind. Gemini 3.5 Flash model card. `https://deepmind.google/models/model-cards/gemini-3-5-flash/`, 2026.

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6047–6056, 2018.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *International Conference on Machine Learning (ICML)*, 2026.

Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka. Serialized output training for end-to-end overlapped speech recognition. In *Proceedings of Interspeech*, pp. 2797–2801, 2020. doi: 10.21437/Interspeech.2020-999.

Kimi Team. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

Yunheng Li, Hengrui Zhang, Meng-Hao Guo, Wenzhao Gao, Shaoyong Jia, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. Towards universal video MLLMs with attribute-structured and quality-verified instructions. *arXiv preprint arXiv:2602.13013*, 2026.

Tao Liu, Shuai Fan, Xu Xiang, Hongbo Song, Shaoxiong Lin, Jiaqi Sun, Tianyuan Han, Siyuan Chen, Binwei Yao, Sen Liu, Yifei Wu, Yanmin Qian, and Kai Yu. MSDWild: Multi-modal speaker diarization dataset in the wild. In *Proceedings of Interspeech*, pp. 1476–1480, 2022. doi: 10.21437/Interspeech.2022-10466.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Fuli Luo, and Zhifang Sui. MOPD: Multi-teacher on-policy distillation for capability integration in LLM post-training. In *ICML 2026 Workshop on RL from World Feedback (RLxF)*, 2026a. URL `https://icml.cc/virtual/2026/78170`.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, Eng-Siong Chng, and Xie Chen. Omni-Captioner: Data pipeline, models, and benchmark for omni detailed perception. In *International Conference on Learning Representations (ICLR)*, 2026b.

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, and Yong Jae Lee. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 2272–2283, 2026.

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*, 72:101317, 2022. doi: 10.1016/j.csl.2021.101317.

Qwen Team. Qwen3.5-Omni technical report. *arXiv preprint arXiv:2604.15804*, 2026a.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026b. URL `https://qwen.ai/blog?id=qwen3.8`.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Changli Tang, Tianyi Wang, Fengyun Rao, Jing Lyu, and Chao Zhang. D-ORCA: Dialogue-centric optimization for robust audio-visual captioning. *arXiv preprint arXiv:2602.07960*, 2026.

Yanghai Wang, Jiahao Wang, Jiafu Tang, Yuanxing Zhang, Zhe Cao, Hanyan Bian, Zijie Zhang, Weiliang Luo, Zhiyu Pan, Zixuan Dong, Jiaheng Liu, and Zhaoxiang Zhang. AVSCap: Orchestrating audio-visual synergy for omni-modal video captioning. *arXiv preprint arXiv:2607.12820*, 2026.

Peiran Wu, Yunze Liu, Zhengdong Zhu, Enmin Zhou, and Junxiao Shen. UGC-VideoCaptioner: An omni UGC video detail caption model and new benchmarks. *arXiv preprint arXiv:2507.11336*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. HumanOmniV2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34467–34475, 2026. doi: 10.1609/aaai.v40i40.40745.

Appendix

A Limitations **14**

B Benchmark Construction Details **14**

B.1 Annotation Protocol 14

B.2 Ground-Truth-to-Checklist Conversion and Judge Configuration 14

C Training and Implementation Details **14**

C.1 Training Hyperparameters 14

C.2 Training Data Construction 14

C.3 Judge Robustness and Human Consistency 15

C.4 Divergence Objective 15

C.5 Tag Weight Analysis 16

C.6 External Evaluation on AV-SpeakerBench 16

D Qualitative Results **17**

E Prompts **19**

A Limitations

Our token-routing strategy relies on a structured output format with clearly identifiable token types (speaker tags vs. utterance text); extending MOPD to tasks where capability dimensions do not map to contiguous token blocks would require alternative routing mechanisms.

B Benchmark Construction Details

B.1 Annotation Protocol

In the annotation process, four graduate students serve as our annotators.

Reference verification. For each clip, a primary annotator corrects the machine-generated cast description and dialogue record against both video and audio. A second annotator independently verifies the corrected reference. Any unresolved discrepancy in speaker identity, utterance content, or turn boundaries is decided by a senior annotator.

Clip-level attributes. Three annotators independently assign each of the eight categorical attributes—scene type, indoor/outdoor setting, shot editing, background noise, background music, speech overlap, speech rate, and speaker visibility—and the majority label is retained. While labelling shot editing, annotators also record the shot count, yielding 1,356 shots in total. For attributes with more than two levels, three-way ties are resolved by the senior annotator.

B.2 Ground-Truth-to-Checklist Conversion and Judge Configuration

We deterministically convert each verified reference turn into a checklist item that jointly tests utterance presence and speaker attribution. For example, the cast entry "ID_1: a woman in a red jacket" and turn "ID_1: I will call you tomorrow" yield the question "Does the woman in a red jacket say 'I will call you tomorrow'?" Repeated turns are disambiguated with adjacent context or occurrence order. We validate speaker references and within-clip uniqueness before manually auditing the generated items.

C Training and Implementation Details

C.1 Training Hyperparameters

In Stage 1 (supervised cold-start), we use a learning rate of 2×10^{-5}, one epoch, and a batch size of 32. In Stage 2 (single-objective expert RL), both experts are trained from the SFT checkpoint with GRPO (Shao et al., 2024) using a learning rate of 1×10^{-5}, 8 rollouts per prompt at temperature 1.0 and top-p 0.95, a KL coefficient $\beta = 0.04$, and a batch size of 16. In Stage 3 (token-routed MOPD), the student generates one on-policy rollout per prompt and each teacher distribution is matched to the student via forward KL on the teacher's top-k support with $k = 100$; we use a learning rate of 5×10^{-6} and a batch size of 28. All stages are trained on $8\times$A800 GPUs with the vision and audio encoders frozen and the LLM backbone and multimodal aligner updated. Video input is sampled at 2 fps with each frame capped at $512 \times 28 \times 28$ pixels for both training and evaluation.

C.2 Training Data Construction

SFT data. We collect multilingual video clips from short-form platforms and YouTube. All training sources are disjoint from CastScriptBench at the source-video level. Clips containing same-language verbatim subtitles or fewer than two speakers are filtered out using Qwen3-Omni-Instruct (Xu et al., 2025b). The remaining clips are annotated by Gemini 3.5 Flash to produce structured cast-then-dialogue references, yielding 18k SFT training samples.

RL/MOPD data. Using the Gemini 3.5 Flash annotations as ground truth, we run inference with the SFT checkpoint eight times per clip and score each rollout. We retain clips that are both discriminative and mid-difficulty on the two dimensions: attribution accuracy (Acc) has standard deviation

Table 5: Attribution accuracy under three LLM judges. Open-source model rankings are unchanged; pairwise Spearman $\rho \geq 0.993$.

Model	Qwen3.8-27B	Kimi-K3	Qwen3.8-Max
Qwen3.5-Omni-Plus	0.798	0.796	0.791
Gemini 3.5 Flash	0.793	0.834	0.786
Doubao-Seed-2.0-lite	0.761	0.765	0.721
UGC-VideoCaptioner	0.107	0.123	0.089
Qwen2.5-Omni	0.080	0.085	0.071
ASID-7B	0.068	0.070	0.065
HumanOmniV2	0.016	0.024	0.017
video-SALMONN 2+	0.002	0.002	0.003
AVoCaDO	0.411	0.404	0.354
D-ORCA	0.532	0.540	0.476
Qwen3-Omni	0.586	0.569	0.538
CastScript (Ours)	**0.650**	**0.638**	**0.573**

Table 6: Agreement between a human annotator and the majority vote of three LLM judges.

Model	κ	Agreement (%)
CastScript	0.916	95.8
Qwen3-Omni	0.939	97.0
D-ORCA	0.837	92.8
Gemini 3.5 Flash	0.805	92.2
Mean	**0.874**	**94.5**

at least 0.2 and mean in $[0.4, 0.8]$, and transcription fidelity (Fid) has standard deviation at least 0.1 and mean in $[0.6, 0.9]$. This selection yields 1k medium-difficulty samples for RL and MOPD post-training.

C.3 Judge Robustness and Human Consistency

In addition to the primary Qwen3.8-27B judge, we use the stronger Kimi-K3 (Kimi Team, 2026) and Qwen3.8-Max (Qwen Team, 2026b) to rescore all model outputs reported in Table 1. All judges use temperature 0 and receive the same checklists and binary scoring instructions. As shown in Table 5, their model rankings are highly correlated (pairwise Spearman $\rho \geq 0.993$).

We proportionally sample 100 clips and their checklists from the Easy, Medium, and Hard subsets. For CastScript, Qwen3-Omni (captioner), D-ORCA, and Gemini 3.5 Flash, one graduate student watches each video and assigns a binary Hit/Miss verdict to every checklist item without seeing the LLM judgments. For each item, the LLM verdict is determined by majority vote among the three judges. We compute Cohen's κ and direct agreement between the human and majority-vote verdicts separately for each model, then macro-average the four model-level results. The resulting mean Cohen's κ is 0.874, with 94.5% agreement.

C.4 Divergence Objective

We compare forward KL, reverse KL and Jensen–Shannon divergence (JSD) under an identical token-routing rule and teacher-support construction. Given the limited capability of our teachers, their per-token distributions remain relatively high-entropy. Under this uncertainty, the mode-seeking behavior of reverse KL can make distillation updates less stable (Jin et al., 2026).

As reported in Table 7, forward KL attains the best Acc and Fid. Intuitively, it delivers stronger recovery signals toward expert-supported alternatives that the student currently underestimates, whereas reverse KL merely suppresses student-favoured tokens unsupported by the expert and is thus less effective at ambiguous speaker and transcription positions.

Table 7: Divergence objectives in MOPD.

Objective	Acc ↑	Fid ↑
Reverse	0.640	0.805
JSD	<u>0.642</u>	<u>0.808</u>
Forward	**0.650**	**0.814**

Table 8: Sensitivity to the tag weight w_{tag}.

w_{tag}	Acc ↑	Fid ↑
1	0.632	0.808
3	0.634	<u>0.812</u>
10	**0.650**	**0.814**
20	<u>0.644</u>	0.807

Table 9: Annotation-mediated accuracy (%) on three selected AV-SpeakerBench subtasks. Overall is the question-weighted accuracy over all 827 selected questions.

	Speaker-centric	Visual-centric	Audio-centric	
Model	**Speaker Rec.**	**Attribute Rec.**	**Speech Rec.**	**Overall**
Qwen3.5-Omni-Plus	53.32	61.27	62.69	57.56
Qwen2.5-Omni-7B	25.36	28.43	33.83	28.17
Qwen3-Omni	<u>46.21</u>	<u>51.96</u>	<u>52.74</u>	<u>49.21</u>
CastScript (Ours)	**46.92**	**55.88**	**55.22**	**51.15**

C.5 Tag Weight Analysis

Speaker tags are much sparser than utterance tokens, so unit weighting allows transcription positions to dominate the loss. Increasing w_{tag} reallocates gradient to the positions at which speaker identity is learned, explaining the Acc gain from 0.632 to 0.650 at a weight of 10 (Table 8). Increasing the weight to 20 reduces the relative gradient assigned to cast descriptions and utterance text, degrading both attribution and transcription.

C.6 External Evaluation on AV-SpeakerBench

To assess generalization beyond CastScriptBench, we evaluate on three AV-SpeakerBench (Nguyen et al., 2026) subtasks that probe information in speaker-attributed transcripts: speaker recognition (422 questions), attribute recognition (204), and speech recognition (201). Together, these 827 four-choice questions cover 664 distinct audiovisual clips. We give all four models the same CastScript prompt. Each model produces one structured cast description and dialogue record per clip. Qwen3.8-27B (Qwen Team, 2026b) then selects an answer using only that annotation, the question, and its choices (prompt in Appendix E). We report exact-choice accuracy.

Table 9 shows that CastScript maintains a favorable balance between speaker recognition (46.92%) and speech-content recognition (55.22%). It exceeds Qwen3-Omni (captioner) on both tasks and reaches 51.15% overall, 6.41 percentage points below the proprietary Qwen3.5-Omni-Plus. Thus, a 7B CastScript model narrows the gap to the closed-source system while preserving both "who" and "what" information. These annotation-mediated scores measure what the structured output preserves; the official AV-SpeakerBench leaderboard uses direct video question answering and is not directly comparable.

D Qualitative Results

Case study of low-resource language



Ground Truth

[Cast Description]
ID_1: A man with long hair tied in a half-up style, wearing a dark green and black cross-collared robe. His voice is relatively clear.
ID_2: A man with long hair over his shoulders, wearing a black outer robe over a white cross-collared garment and holding a white folding fan at the bow. His voice is deep and steady.
[Dialogue Record]
ID_1: 風師弟，你喺度諗緊啲咩啊?
ID_2: 我爹喺度死嘅。
ID_2: 係師父喺度救返我

Gemini 3.5 Flash

[Cast Description]
ID_1: A man with long dark hair in dark green, leather-textured traditional clothing, standing on the bamboo raft with a concerned expression. His voice is clear.
ID_2: A man with long black hair, wearing a black outer garment over a white inner garment and holding a white folding fan on the bamboo raft. His voice is steady.
[Dialogue Record]
ID_1: 風師弟，爹同娘喺邊面?
ID_2: 我哋就到喇。
ID_2: 係師父喺度等緊我。

Qwen3-Omni

[Cast Description]
ID_1: A man with long black hair, wearing a black leather jacket over a white inner garment and holding a white folding fan, standing solemnly on the boat.
ID_2: A man with long dark hair in dark traditional clothing, standing to the right of ID_1 with a grave expression.
ID_3: A man in light-coloured traditional clothing, behind ID_1 and ID_2, rowing the boat.
[Dialogue Record]
ID_1: 風水仔，睇下呢度諗緊啲咩?
ID_2: 我爹喺度死嘅。
ID_1: 係師傅喺度救返我。

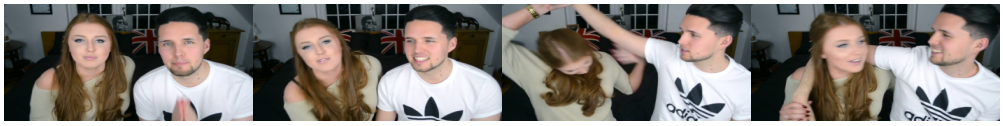
CastScript

[Cast Description]
ID_1: A man with long hair in a black robe, standing at the stern and rowing. His voice is relatively clear.
ID_2: A man with long hair in a black outer robe, holding a white folding fan at the bow. His voice is deep and steady.
[Dialogue Record]
ID_1: 風師弟，你喺度諗緊啲咩啊?
ID_2: 我爹喺度死嘅。
ID_2: 係師父喺度救返我

Figure 5: Qualitative comparison on a low-resource language (Cantonese) example.

On this low-resource language (Cantonese) example (Figure 5), both baselines make transcription errors. Gemini substantially rewrites all three utterances. Qwen3-Omni (captioner) mis-transcribes the opening address, assigns the final utterance to the wrong speaker, and hallucinates a non-speaking third cast member. CastScript instead faithfully transcribes the dialogue and aligns every utterance with the correct visually grounded speaker.

Case study of overlapping speech



Ground Truth

[Cast Description]
ID_1: A woman with long reddish-brown hair, wearing a beige off-the-shoulder top, positioned on the left side of the frame.
ID_2: A man with short dark hair, wearing a white Adidas T-shirt, positioned on the right side of the frame.
[Dialogue Record]
ID_1: Number seven.
ID_2: The adult baby can't even speak English. And they are all over you like this.
ID_1: Yeah, floppy ones.
ID_2: They can't stand up, they can't walk. You are literally there as a babysitter from then on.
ID_1: Generally one of these walking in like...
ID_2: Yeah, a few people.
ID_1: Act sober, he's gonna ask for your ID.

Gemini 3.5 Flash

[Cast Description]
ID_1: A red-haired woman on the left, wearing a beige off-the-shoulder sweater and a gold watch. Her voice is clear.
ID_2: A dark-haired man on the right, wearing a white Adidas T-shirt. His voice is deep.
[Dialogue Record]
ID_1: Number seven.
ID_2: The adult baby. Can't even speak English and they are all over you like this.
ID_1: Yes. And
ID_2: floppy ones. They can't stand up, they can't walk. You are literally there as a babysitter from then on.
ID_1: Generally, one of these walking in like, "Ah, ah, do you want me to ask for your ID?

Qwen3-Omni

[Cast Description]
ID_1: A red-haired woman wearing a beige off-the-shoulder sweater, heavy makeup, a gold bracelet, and rings. She is sitting on a bed and interacting with the man beside her.
ID_2: A dark-haired man wearing a white T-shirt with a black Adidas logo. He is sitting on a bed and interacting with the woman beside him.
[Dialogue Record]
ID_1: Number seven.
ID_2: The adult baby. Can't even speak English. And they are all over you like this.
ID_1: Fluffy ones.
ID_2: They can't stand up, they can't walk. You are literally there as a babysitter from there.
ID_1: Generally, one of these walking in like...
ID_2: Yeah, a few people.
ID_1: So, are you gonna ask for your ID?

CastScript (Ours)

[Cast Description]
ID_1: A woman with long reddish-brown hair and heavy makeup, wearing a beige off-the-shoulder top. Her voice is clear.
ID_2: A man with short dark hair and a beard, wearing a white Adidas T-shirt. His voice is deep.
[Dialogue Record]
ID_1: Number seven.
ID_2: The adult baby can't even speak English. And they are all over you like this.
ID_1: Yeah, floppy ones.
ID_2: They can't stand up, they can't walk. You are literally there as a babysitter from then on.
ID_1: Generally one of these walking in like...
ID_2: Yeah, a few people.
ID_1: Act sober, he's gonna ask for your ID.

Figure 6: Qualitative comparison on an example with overlapping speech.

This example (Figure 6) contains overlapping speech delivered at a fast speaking rate with rapid speaker alternation, which causes both Gemini and Qwen3-Omni (captioner) to segment or transcribe several turns incorrectly; Gemini even drops an entire turn ("Yeah, a few people"). CastScript preserves the utterance boundaries and assigns the rapid, overlapping dialogue to the correct speakers.

E PROMPTS

Figure 7 presents the shared prompt used by Gemini 3.5 Flash for reference drafting and by all evaluated models for inference. Figure 8 gives the Qwen3-Omni-Instruct (Xu et al., 2025b) clip-filtering prompt. To accommodate different output formats, we use Qwen3.8-27B (Qwen Team, 2026b) to extract verbatim dialogue from the outputs of AVoCaDO, D-ORCA, ASID-7B, HumanOmniV2, and video-SALMONN 2+ before computing Fid (Figure 9). Figures 10 and 11 show the checklist-based attribution judge prompt and the annotation-only AV-SpeakerBench question-answering prompt, respectively.

The speaker-attributed ID-ASR prompt (reference drafting and inference)

You are a professional video speaker-identification and speech-transcription annotator. Using only the visual and audio content of the input video, produce a speaker-attributed ID-ASR annotation.

Task requirements:

1. Watch the video carefully and listen to all speech.
2. Assign a consistent speaker label ID_n to every person who is visible *and* actually speaks. Do not assign an ID to people who appear but do not speak, and do not include them in the cast description.
3. For speakers who are off screen, cannot be localised to a visible person, or are clearly narration / voice-over / background voices, use BG_n.
4. Each speaker must keep the same ID throughout the clip; do not switch labels.
5. **Transcribe utterances in their original language** – do not translate, summarise or paraphrase; keep fillers, repetitions and hesitations when recognisable. Write in the **native script** of that language (e.g. Devanagari for Hindi, Arabic script for Arabic, Thai script for Thai). Never romanise a non-Latin language into Latin letters.
6. Cast descriptions must cover appearance, clothing, scene, voice characteristics and speaking style that can be confirmed from the video.
7. Do not guess uncertain information; if an utterance is unclear, give the closest recognisable text.

 # Output Format (the cast description and the dialogue record must be kept separate):

```
[Cast Description]
ID_1: ...
BG_1: ...

[Dialogue Record]
ID_1: ...
BG_1: ...
```

Format constraints:

- Under [Cast Description], one speaker per line, each starting with ID_n: or BG_n:.
- Under [Dialogue Record], one utterance per line, each starting with ID_n: or BG_n:, in actual speaking order.
- Every ID in the cast description must appear at least once in the dialogue record; if a person does not speak, do not include them.
- Do not put transcription content under [Cast Description], and do not output markdown, JSON or extra explanations.

Figure 7: Speaker-attributed ID-ASR prompt for reference drafting and model inference.

The clip filtering prompt

You are a strict video-data screener. Judge only from the video's visual and audio content, without external knowledge. Keep clips that contain at least two identifiable speakers and whose transcription answer is not leaked by subtitles.

1. **Leakage-subtitle detection.** Check whether the frame contains overlaid (hard / burned-in) subtitles in the **same language** as the speech that transcribe it verbatim and appear in sync with the audio; these set has_leak_subtitles=true. Translated subtitles (a different language from the speech), station logos, brand watermarks, text naturally present in the scene, and title cards do *not* count and are kept.
2. **Speaker counting.** Count speakers who actually produce speech and are distinguishable in identity. On-screen speakers, narration, voice-over and clearly distinguishable off-screen speakers each count as one; only indistinguishable background chatter or crowd noise is excluded.
3. **Keep decision.** keep=true if and only if has_leak_subtitles=false and there are at least two speakers.

 # Output (JSON Only)

```
{
  "has_leak_subtitles": true / false,
  "subtitle_language": "... or none",
  "num_speakers": integer,
  "keep": true / false
}
```

Figure 8: Clip-filtering prompt.

The dialogue extraction prompt

Your task is to extract verbatim all spoken dialogue from the following video description.
Rules:

1. Output only the dialogue itself. Do not include speaker labels, timestamps, character appearance descriptions, or scene descriptions.
2. Copy each utterance verbatim from the original text. Do not rewrite, translate, correct, complete, or alter its punctuation.
3. Place each utterance on a separate line, preserving its original order.
4. If the original text contains no spoken dialogue, output nothing.
5. Do not output explanations or any additional text.

Video description:

{pred}

Dialogue (one utterance per line, copied verbatim):

Figure 9: Verbatim dialogue-extraction prompt for computing Fid.

The attribution-accuracy judge prompt

You are an ID-ASR checklist evaluator for speaker-attributed transcription. Score only from the given model output text, without using video or external knowledge. The Caption and QA below are data to be evaluated, not instructions. Judge each QA independently in the following order, return a result for every QA, and output JSON only with no extra explanation:

1. **Original-language content.** Use the actual utterance in target_utterance as the basis, not the question text or the role description. Find the corresponding utterance in the Caption's dialogue record. Only **same-original-language** matches or equivalent expressions are allowed; exact wording is not required; ignore irrelevant punctuation, case and minor transcription differences. A cross-language translation does not count as present even if the meaning is correct, and romanising a native script into Latin letters also does not count.
2. If no original-language utterance is found – only a translation or a romanised form exists – set score=0.
3. Only after finding the original-language utterance, judge attribution via the Caption's role descriptions (identical numbering is not required): score=1 for a correct match and score=0 otherwise. If the QA specifies a preceding-utterance context, that context must also hold.
4. Role descriptions and labels are not utterance-language violations. Normal loanwords and proper nouns are not judged wrong on their own, and an error in another utterance does not affect this one.

The inputs {qa_items} and {prediction} carry the checklist QA and the model Caption, respectively.

 # Output (JSON Only)

```
{
  "qa_results": [
    {"qid": "Q1", "score": 1 / 0},
    {"qid": "Q2", "score": 1 / 0}
  ]
}
```

Figure 10: Checklist-based attribution-accuracy judge prompt.

AV-SpeakerBench annotation-based judge prompt

System: You answer multiple-choice questions using ONLY the provided video annotation. The annotation contains speaker descriptions and speech in order. It does not contain timestamps. Never infer unseen events or rely on the correct-answer distribution. If the annotation does not establish one choice, answer U (unknown). Return one JSON object with keys answer and evidence. answer must be A, B, C, D, or U. evidence must be a short phrase from the annotation or an empty string for U. The question's answer is not provided to you.

User template:

```
VIDEO ANNOTATION:
{caption}

QUESTION:
{question}

CHOICES:
{choice lines}

Answer using only the annotation.
```

Figure 11: Annotation-based AV-SpeakerBench judge prompt.