AUTOCAST: AUTOMATED CHARACTER-ATTRIBUTED SPEECH TRANSCRIPTION IN LONG-FORM TV DRAMAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated named-speaker transcription aims to determine the start and end times, text, and speaking character of each utterance in long-form TV dramas. This task remains challenging for existing audio-visual models: multiple visible characters, off-screen speech, and reaction shots create visual ambiguity, while the lack of reliable character-specific voice references and limited speaker discrimination complicate acoustic attribution. To address these challenges, we propose **AutoCAST** (Automated Character-Attributed Speech Transcription), a three-stage framework that takes raw audiovisual streams, a cast roster, and character portraits as input to produce timestamped, character-attributed transcripts. Timestamp-aware transcription first predicts utterance boundaries and text using explicit temporal markers. Agreement-guided reference bank construction then selects candidate speech segments whose initial identity predictions agree with face observations to build a character-specific voice reference bank. Finally, reference-guided multimodal attribution performs acoustic seed filtering and label propagation, then combines character-level voiceprint scores with visual and dialogue context to determine the speaker of each utterance. The complete framework requires neither manual transcripts nor manually verified voice references at inference. On DramaSR-532K, AutoCAST achieves 91.70% mean speaker-attribution accuracy across seven held-out Chinese-language dramas, outperforming the strongest evaluated closed-source baseline (83.06%) by 8.64 percentage points, and achieves 87.27% across three held-out English-language series. Ablation studies support the effectiveness of automatic segmentation, reference bank construction, and multimodal speaker attribution. Code is available at `https://anonymous.4open.science/r/AutoCAST/`.

1 INTRODUCTION

Recent advances in vision-language models and video captioners have improved general video understanding, with strong performance on broad multimodal benchmarks (Zhang et al., 2024; Li et al., 2024; Zhang et al., 2025; Zhu et al., 2025; Chai et al., 2025; Tang et al., 2025; Chen et al., 2026). However, audio-visual models still face substantial difficulties in identifying *who says what and when* (Nguyen et al., 2026; Tang et al., 2026). The meaning of an utterance depends on the speaker and its relationship to preceding dialogue, making character identity central to story understanding (Raajesh et al., 2024), while temporally grounded dialogue provides essential context for video question answering (Lei et al., 2018). Long-form television dramas impose a further requirement: each speaking character must retain a consistent name across scenes and episodes, allowing their dialogue to connect narrative events throughout a series.

In this paper, we study **automated named-speaker transcription**: given raw video and audio, a cast roster, and reference character portraits, the system outputs the start and end timestamps, transcript, and corresponding character name for each utterance. Prior work demonstrates the value of character-specific voice references by combining voiceprint matching with audio-visual context for named-speaker attribution (Li et al., 2026). Such reference-based attribution involves extracting single-speaker speech segments, linking them to character names to construct a reference bank, and using these references with audio-visual context for speaker attribution. However, perform-

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.

ing these steps without manual transcripts or identity-verified voice references presents three challenges: (i) inaccurate utterance boundaries can mix speech from different speakers, compromising voiceprint extraction and reference quality; (ii) automatically assigned reference labels may be incorrect, spreading identity errors across scenes when these references are reused; (iii) even reliable references may be difficult to exploit through direct audio input, as speech-aware large language models remain limited in speaker verification (Thebaud et al., 2026).

To address these three challenges, we propose **AutoCAST** (Automated Character-Attributed Speech Transcription), a three-stage pipeline for named-speaker transcription in long-form TV dramas. First, to reduce contamination from neighboring speakers, **Timestamp-Aware Transcription** interleaves integer-second text markers with raw audio segments before encoding to guide boundary prediction while jointly generating utterance transcripts. The predicted boundaries define the speech crops used for voiceprint extraction and reference bank construction. Next, to improve reference label reliability and limit error propagation, **Agreement-Guided Reference Bank Construction** uses a learned multimodal model to propose character labels for speech segments, then filters these candidates by prediction–face agreement. This screening selects speech examples whose predicted identities are supported by visual observations for subsequent acoustic matching. Finally, to address the difficulty of using reference audio directly, **Reference-Guided Multimodal Attribution** applies acoustic seed filtering and label propagation to derive ranked character candidates and voiceprint scores for each utterance. A multimodal attribution model combines this explicit acoustic evidence with video and dialogue context to determine the final identity. The complete inference pipeline requires neither manual transcripts nor identity-verified voice references.

We evaluate AutoCAST on seven Chinese-language dramas and three English-language series from DramaSR-532K (Li et al., 2026), all excluded from task-specific training. Using the automatic pipeline, AutoCAST achieves mean speaker-attribution accuracies of 91.70% and 87.27%, respectively, outperforming the strongest evaluated closed-source baseline in each language group by 8.64 and 3.53 percentage points. Ablations show that automatic segmentation improves transcription performance and that identity initialization and prediction–face agreement improve reference quality. They further show that multimodal attribution improves over acoustic propagation, with complementary contributions from acoustic and visual evidence.

2 Related Work

2.1 Speaker Recognition and Speaker Diarization

Speaker recognition compares embeddings such as x-vectors and ECAPA-TDNN against enrolled speech (Snyder et al., 2018; Desplanques et al., 2020), whereas speaker diarization determines who speaks when through neural activity prediction or segmentation and clustering (Fujita et al., 2019; Horiguchi et al., 2022; Bredin, 2023). WhisperX improves transcription timing through forced alignment (Bain et al., 2023), and DiarizationLM refines speaker assignments through textual post-processing (Wang et al., 2024a). SpeakerLM unifies ASR and diarization with flexible speaker registration (Yin et al., 2026); TagSpeech uses interleaved temporal anchors for speaker–content alignment (Huo et al., 2026). MOSS and VibeVoice-ASR jointly transcribe and diarize long audio (AI et al., 2026; Peng et al., 2026). Streaming Sortformer and JEDIS-LLM maintain cross-chunk consistency using acoustic or speaker-prompt caches (Medennikov et al., 2025; Shi et al., 2026). AutoCAST additionally grounds acoustic identities in character names using automatically acquired audio-visual references.

2.2 Audio-Visual Reasoning and Character Identification

Active speaker detection associates speech with visible people. AVA-ActiveSpeaker provides an in-the-wild benchmark (Roth et al., 2020), while TalkNet and LoCoNet model temporal and cross-modal context to predict speaking status (Tao et al., 2021; Wang et al., 2024b). Earlier character identification combines cast-list supervision, face context, and voice models (Nagrani & Zisserman, 2017). AutoCAST uses prediction–face agreement for reference admission and reference-conditioned multimodal attribution.

video-SALMONN and video-SALMONN-o1 develop joint audio-visual understanding and reasoning (Sun et al., 2024; 2025); video-SALMONN 2 optimizes caption completeness and ac-

curacy (Tang et al., 2025). AMUSE and HumanOmni-Speaker study multi-speaker understanding (Chowdhury et al., 2026; Bai et al., 2026). AVoCaDO optimizes temporal coherence and dialogue accuracy (Chen et al., 2026), while D-ORCA explicitly optimizes speaker attribution, transcription, and temporal grounding (Tang et al., 2026). AutoCAST emphasizes cross-scene named identity consistency through automatically acquired reference banks. DramaSR-LRM combines voiceprints and scene context for drama recognition, but requires aligned dialogue and verified acoustic seeds (Li et al., 2026).

2.3 General-Purpose Multimodal Models

VideoLLaMA 2 and Qwen3-Omni integrate audio and visual information for multimodal understanding (Cheng et al., 2024; Xu et al., 2025). MiniCPM-o 4.5, Gemini, and Seed extend multimodal interaction and understanding (Yao et al., 2024; Kavukcuoglu et al., 2026; ByteDance Seed, 2026). Omni-Captioner targets detailed audio-visual descriptions with reduced hallucination (Ma et al., 2026). However, speech-aware language models still face challenges in speaker verification (Thebaud et al., 2026), and AV-SpeakerBench exposes difficulties in jointly identifying who said what and when, especially with more visible speakers (Nguyen et al., 2026). AutoCAST grounds named identities through acoustic reference banks and local visual and dialogue context.

3 Method

AutoCAST transcribes long-form video into timestamped utterances and attributes each utterance to a named character. We first obtain character-annotated video $\widetilde{V}$ from raw video, with face annotations linked to a cast roster $\mathcal{C} = \{c_k\}_{k=1}^{K}$. Preprocessing details are provided in Appendix C.1. AutoCAST then takes $\widetilde{V}$, the corresponding audio A, and $\mathcal{C}$ as input, uses $\widetilde{V}$ as visual identity evidence, and outputs a chronologically ordered sequence of utterance records with character identities:

$$f_{\text{AutoCAST}}(\widetilde{V}, A, \mathcal{C}) = \left((\tau_i^{\text{start}}, \tau_i^{\text{end}}, y_i, \widehat{z}_i)\right)_{i=1}^{N}. \quad (1)$$

Here, N is the number of predicted utterances, and each record contains global start and end times $(\tau_i^{\text{start}}, \tau_i^{\text{end}})$, transcript y_i, and predicted character identity $\widehat{z}_i$. Inference requires neither manual transcripts nor identity-verified reference recordings.



Ground Truth: Chandler
Are there no conscious men in the city for you two?

Doubao's Prediction: Ross ✘
Ross stands in the apartment holding a mug, his mouth moves as he speaks, the labeled Dr Ross Geller matches, and the voice is Ross's distinct speech.

Gemini's Prediction: Ross ✘
Ross is standing in the kitchen speaking.

Figure 1: **Speaker attribution ambiguity.** Both models misattribute Chandler's utterance to Ross despite labeled faces.

Jointly predicting utterance timestamps, transcripts, and speaker identities remains challenging for audio-visual models (Tang et al., 2026; Nguyen et al., 2026), as our evaluation on long-form TV dramas also shows (Table 1). Multiple visible characters, off-screen speakers, and reaction shots make face identity alone insufficient for attribution (Figure 1), motivating character-specific acoustic references.

We propose AutoCAST, a three-stage framework (Figure 2). First, Timestamp-Aware Transcription (Section 3.1) uses explicit temporal markers to jointly predict utterance boundaries and text for subsequent voiceprint extraction. Second, Agreement-Guided Reference Bank Construction (Section 3.2) builds character-specific voice references by selecting speech samples whose predicted identities agree with face observations. Third, Reference-Guided Multimodal Attribution (Section 3.3) addresses speech-aware language models' limited speaker-verification ability (Thebaud et al., 2026) by converting reference speech into candidate identities and voiceprint similarity scores, then combining this evidence with visual information and dialogue context. AutoCAST requires neither manual transcripts nor manually verified voice references at inference time.

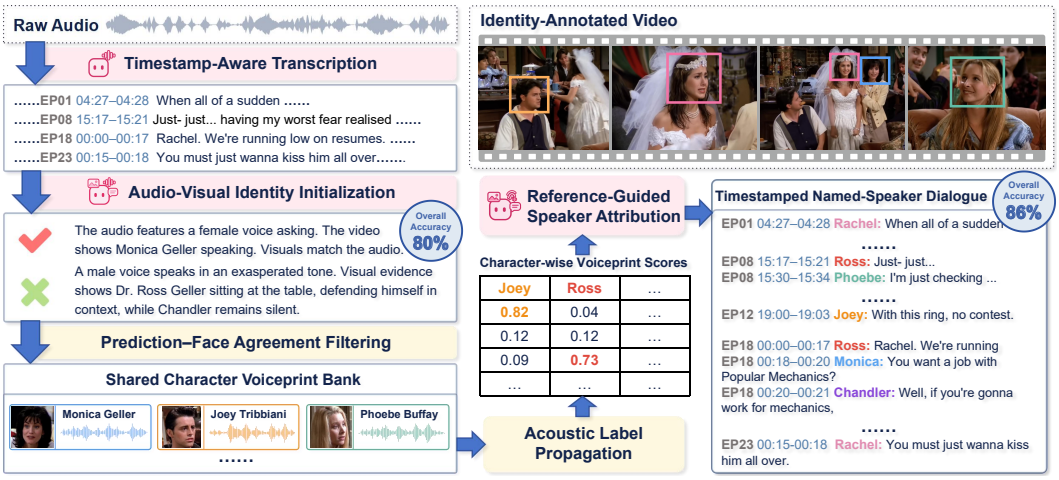


Figure 2: **Overview of the AutoCAST framework.** (1) Timestamp-aware transcription predicts utterance boundaries and text. (2) Agreement-guided bank construction collects character-linked references. (3) Multimodal attribution combines voiceprint scores, visual information, and dialogue context to identify speakers.

3.1 Stage 1: Timestamp-Aware Transcription

We first jointly predict utterance boundaries and text from raw audio, providing the temporal and textual basis for speaker attribution. Accurate boundaries are essential: audio crops that include adjacent speakers or omit target speech can compromise the speaker embeddings used to construct the reference bank. We therefore insert explicit temporal markers into the audio input to guide boundary prediction. Local inference uses windows of at most one minute; window processing and timestamp conversion are described in Appendix C.1. For the raw audio A_b of clip b, let $A_{b,j}$ denote the raw audio segment for the local time interval $[j, j+1)$, truncated at the clip end if necessary. Before audio encoding, we place the integer-second text marker t_j, indicating local time j seconds, before each segment in the multimodal input sequence:

$$A_b^{\text{ts}} = [t_0, A_{b,0}, t_1, A_{b,1}, \ldots]. \tag{2}$$

This sequence is then tokenized and audio-encoded for model input. When constructing training data, we segment utterances using speaker labels and semantic information to account for both speaker changes and semantic completeness. The start time, end time, and text of each utterance form a supervision record:

$$u_i = (\tau_i^{\text{start}}, \tau_i^{\text{end}}, y_i). \tag{3}$$

Inference outputs follow the same structure, with each field predicted by the model. At inference time, the model jointly predicts utterance start times, end times, and text from the timestamp-aware audio input A_b^{ts} and a transcription instruction, generating a local utterance sequence $\mathcal{U}_b^{\text{raw}}$ in speaking order:

$$\mathcal{U}_b^{\text{raw}} = f_{\text{trans}}(A_b^{\text{ts}}). \tag{4}$$

The records retained under the window-processing rule form $\mathcal{U}_b$. We convert their timestamps to global time and merge the records chronologically to obtain the complete utterance sequence $\mathcal{U} = (u_i)_{i=1}^N$. The resulting boundaries define audio crops and face observation windows for reference bank construction, while the transcripts supply dialogue context for identity attribution.

3.2 Stage 2: Agreement-Guided Reference Bank Construction

Using the utterance boundaries and transcripts from Stage 1, this stage collects candidate voice references for each character across the drama. An audio-visual model first predicts an identity for each utterance; prediction–face agreement then selects utterances to populate a shared voice reference bank $\mathcal{Q} = \{\mathcal{Q}_k\}_{k=1}^K$, where $\mathcal{Q}_k$ contains the candidate reference utterances for character c_k. Stage 3 uses the corresponding audio crops and character labels for acoustic label propagation and voiceprint scoring.

3.2.1 AUDIO-VISUAL IDENTITY INITIALIZATION

For unit b, the audio-visual identity initialization model (Init) predicts a character identity for each utterance from the annotated video, timestamp-aware audio, and Stage 1 utterance sequence augmented with up to eight preceding and eight following utterances, denoted by $\mathcal{U}_b^{\mathrm{ctx}}$. Context utterances provide evidence only. The audio and video cover the input window associated with unit b; predictions are requested for its N_b retained target utterances:

$$(\widetilde{z}_{b,i})_{i=1}^{N_b} = f_{\mathrm{attr}}\Big([P_{\mathrm{attr}}; \widetilde{V}_b; A_b^{\mathrm{ts}}; \mathcal{U}_b^{\mathrm{ctx}}]\Big)\,. \tag{5}$$

Here, $N_b = |\mathcal{U}_b|$ is the number of retained target utterances in unit b, $\widetilde{z}_{b,i} \in \mathcal{C}$ is the initial identity prediction for its i-th utterance, and P_{attr} is the attribution instruction. The model combines visual identity cues, acoustic speaker cues, and textual dialogue context. We reindex the predictions across units as $(\widetilde{z}_i)_{i=1}^{N}$, aligned with $\mathcal{U}$, for subsequent agreement filtering.

3.2.2 PREDICTION–FACE AGREEMENT FILTERING

Incorrect reference labels can propagate errors to other utterances during acoustic matching. We therefore retain an utterance predicted as character c_k as a reference candidate only if the same character is observed nearby in time. Let $W_i^{(r)}$ denote the temporal observation window for utterance u_i, where $r \in \{\mathrm{narrow}, \mathrm{wide}\}$ specifies a narrow or wide window. Let $\mathcal{H}(W)$ be the set of characters identified within window W from the character annotations in $\widetilde{V}$. For each character $c_k \in \mathcal{C}$, we filter the utterance collection $\mathcal{U}$ of the entire drama to define:

$$\mathcal{Q}_k = \begin{cases} \Big\{u_i \in \mathcal{U} \mid \widetilde{z}_i = c_k,\ c_k \in \mathcal{H}(W_i^{(\mathrm{narrow})})\Big\}, & \text{if this set is nonempty,} \\ \Big\{u_i \in \mathcal{U} \mid \widetilde{z}_i = c_k,\ c_k \in \mathcal{H}(W_i^{(\mathrm{wide})})\Big\}, & \text{otherwise.} \end{cases} \tag{6}$$

The wide-window set is used only when no narrow-window candidates exist for that character anywhere in the drama; the two sets are not merged. Together, these character-specific sets form the reference bank $\mathcal{Q} = \{\mathcal{Q}_k\}_{k=1}^{K}$. This criterion retains utterances whose predicted identities agree with visual observations.

3.3 STAGE 3: REFERENCE-GUIDED MULTIMODAL ATTRIBUTION

This stage connects the drama-level reference bank to utterance-level multimodal attribution by representing acoustic identity evidence as ranked character candidates and voiceprint similarity scores. Acoustic seed selection and label propagation extend the evidence available for matching across the drama. The final attribution model jointly interprets character-level voiceprint scores, visual observations, and dialogue context to resolve ambiguous speaker assignments.

3.3.1 ACOUSTIC LABEL PROPAGATION

For acoustic processing, each utterance record is associated with the audio segment cropped using its Stage 1 predicted start and end times. Given the reference bank $\mathcal{Q}$ from Stage 2, we follow DramaSR-LRM (Li et al., 2026) to first select acoustic seeds through clustering of the reference samples. Label propagation then assigns utterances to characters based on voiceprint similarity, adding accepted utterances to the corresponding seed sets. These expanded seed sets support subsequent matching, allowing identity evidence to propagate beyond the initial seeds. The procedure returns the initial acoustic seed sets together with the resulting character rankings and voiceprint scores:

$$(\mathcal{S}, \mathcal{K}) = f_{\mathrm{DramaSR}}(\mathcal{U}, \mathcal{Q})\,. \tag{7}$$

Here, $\mathcal{S} = \{\mathcal{S}_k^{\mathrm{seed}}\}_{k=1}^{K}$ contains the character-specific seeds selected before propagation, which are used to evaluate seed quality in the reference bank ablations. Utterances added during propagation are excluded from $\mathcal{S}$. $\mathcal{K} = (\mathbf{k}_i)_{i=1}^{N}$ contains the ranked character candidates and voiceprint scores for each utterance. For a character with $m_k \geq 1$ references in its current set, the backend averages the top $\max(1, \min(50, \lfloor m_k^{0.4} \rfloor))$ cosine similarities between the target speech embedding and those references. Characters without references receive no acoustic score. For unit b, we reindex the corresponding reports locally as $\mathcal{K}_b = (\mathbf{k}_{b,i})_{i=1}^{N_b}$, where $\mathbf{k}_{b,i}$ corresponds to the i-th utterance in $\mathcal{U}_b$. These reports provide acoustic evidence for subsequent multimodal attribution.

3.3.2 REFERENCE-GUIDED MULTIMODAL ATTRIBUTION

For final attribution, we augment the target utterance sequence $\mathcal{U}_b$ with up to eight preceding and eight following utterances from the drama-level sequence $\mathcal{U}$, denoting the expanded input by $\mathcal{U}_b^{\mathrm{ctx}}$. The surrounding utterances provide context only; the prediction targets remain the N_b utterances in $\mathcal{U}_b$. Attribution uses three complementary sources of evidence: face annotations in $\widetilde{V}_b$ associate visible faces with character names; $\mathcal{K}_b$ links each utterance to candidate characters through voiceprint similarity scores and rankings; and $\mathcal{U}_b^{\mathrm{ctx}}$ provides textual dialogue context for interpreting the target utterances. The final model combines these visual, acoustic, and contextual cues. For unit b, attribution is expressed as:

$$(\widehat{z}_{b,i})_{i=1}^{N_b} = f_{\mathrm{final}}\Big([P_{\mathrm{final}}; \widetilde{V}_b; \mathcal{U}_b^{\mathrm{ctx}}; \mathcal{K}_b]\Big). \tag{8}$$

Here, $\widehat{z}_{b,i} \in \mathcal{C}$ is the predicted character identity for the i-th target utterance in unit b. The final attribution instruction P_{final} asks the model to combine this evidence and output exactly one identity for each target utterance. We merge these unit-level predictions chronologically and reindex them as $(\widehat{z}_i)_{i=1}^{N}$. Pairing them with the global boundaries and transcripts from Stage 1 yields the final named-speaker transcript, including utterances not selected as reference samples.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data and supervision. We conduct our experiments on DramaSR-532K (Li et al., 2026), which contains 13 TV series, 806 videos, and approximately 532K annotated dialogue lines over 525 hours and 31 minutes. Following the train–test split of DramaSR-532K, we use only the two Chinese-language dramas *A Lifelong Journey* and *Empresses in the Palace* for task-specific training; no English-language series is included in this training set. The Chinese-language evaluation covers seven dramas: *The Long Night*, *Qin Empire 2*, *Ode to Joy 1*, *The Knockout*, *Minning Town*, *Stand by Me 1*, and *Battle of Changsha*. The English-language evaluation covers *Friends*, *Lost*, and *Downton Abbey* to assess cross-language generalization. No evaluation series is used for task-specific training. Table 1 reports complete-pipeline results for both language groups; the ablation studies use the seven Chinese-language dramas. Detailed results are provided in Appendix B.

Metrics. We evaluate transcription using character error rate (CER) for Chinese-language dramas and word error rate (WER) for English-language series, together with mean absolute boundary error (AAS) and Single-Speaker GT Coverage (SSGC). AAS measures the mean absolute error of utterance start and end boundaries in seconds: boundary errors can truncate target speech or include neighboring speakers, compromising voiceprint extraction and reference quality. SSGC measures the proportion of GT utterances covered by at least one matched prediction associated with a single GT speaker, irrespective of predicted character names. Let $\mathcal{U}_{\mathrm{GT}}$ denote all evaluated GT utterances and $\mathcal{G}_p$ the GT utterances matched to prediction p. Define $\mathcal{P}_{\mathrm{ss}}$ as the matched predictions whose associated GT utterances share one speaker identity. Then

$$\mathrm{SSGC} = \frac{|\mathcal{C}_{\mathrm{ss}}|}{|\mathcal{U}_{\mathrm{GT}}|}, \qquad \mathcal{C}_{\mathrm{ss}} = \bigcup_{p \in \mathcal{P}_{\mathrm{ss}}} \mathcal{G}_p. \tag{9}$$

Together, AAS and SSGC complement text error rates by measuring boundary precision and single-speaker coverage. For reference bank construction, we evaluate the identity-label accuracy of initial seeds retained before propagation using character-macro seed quality (SeedQual). Let C be the complete evaluation character set for a drama. For construction method m, let $n_{m,c}$ denote the number of seeds assigned to character c, of which $k_{m,c}$ have correct identity labels. We define

$$\mathrm{SeedQual}_m = \frac{1}{|C|} \sum_{c \in C} P_{m,c} \times 100\%, \qquad P_{m,c} = \frac{k_{m,c}}{\max(1, n_{m,c})}. \tag{10}$$

Each character contributes equally to the average, with seed precision set to zero for characters without seeds. Speaker-attribution accuracy (Acc.) is the percentage of GT utterances assigned the correct character identity, with each GT utterance counted once. A GT utterance is scored as correct

only if it has at least one matched prediction and all associated predictions have correct identity labels. Unmatched utterances and those with any incorrect, empty, or invalid predicted identity are scored as incorrect. Reported metrics are averaged equally across the evaluation series within each language group. Metric-specific evaluation units, definitions, alignment rules, and calculation details are given in Appendix A.

Training. We separately fine-tune the transcription, identity initialization, and final attribution models. The transcription model is supervised with GT utterance boundaries and text. Init and final attribution use GT boundaries, transcripts, and identities, with structured rationale-and-answer targets. All models minimize autoregressive cross-entropy over all output tokens. For Init and final attribution, these tokens include both rationales and identity answers. Only the language branch is updated; visual and audio encoders and multimodal projectors remain frozen. Each model is trained for three epochs on eight A800 80GB GPUs using BF16 and `paged_adamw_8bit`, with a learning rate of 10^{-5}, an effective batch size of 32, and cosine decay with 10% warmup. Visual inputs are sampled at 1 FPS during both training and inference.

Inference settings. Prediction–face agreement requires a matching character label with face-recognition confidence of at least 0.45 within the utterance interval extended by one second on each side. If a character has no qualifying candidate, the extension is increased to 20 seconds on each side. The complete offline pipeline uses Stage 1 predicted boundaries and transcripts to construct reference banks and attribute utterances. References are pooled across the entire drama, including future clips. Final attribution receives video frames, predicted utterance records, voiceprint scores, and up to eight utterances preceding and following the target sequence. We use the final-epoch checkpoints and greedy decoding at temperature 0, extracting identity answers from the generated rationale-and-answer outputs. Supervision formats and prompts are provided in the appendix.

4.2 Main Results

We evaluate end-to-end named-speaker transcription against the open-source models Qwen3-Omni (Xu et al., 2025) and MiniCPM-o 4.5 (Yao et al., 2024); the closed-source baselines are Gemini 3.5 Flash (Kavukcuoglu et al., 2026) and Doubao-Seed-2.0-lite (ByteDance Seed, 2026). All methods receive character-annotated video, original audio, and the same cast roster, without manual transcripts or manually labeled voice references.

Table 1: **End-to-end named-speaker transcription performance on Chinese and English DramaSR-532K videos.** All metrics are percentages except AAS (seconds).

Model	DramaSR-532K (Zh)				DramaSR-532K (En)			
	Acc.↑	CER↓	SSGC↑	AAS↓	Acc.↑	WER↓	SSGC↑	AAS↓
MiniCPM-o 4.5	49.72	37.94	62.65	6.3198	33.19	64.93	35.56	14.2721
Qwen3-Omni	62.83	22.75	78.87	4.1381	57.77	24.52	90.78	8.7212
Gemini 3.5 Flash	74.77	24.54	78.70	1.0043	83.74	14.93	91.82	4.2322
Doubao-Seed-2.0-lite	83.06	13.30	87.29	2.6579	79.56	**12.78**	91.46	1.9244
AutoCAST	**91.70**	**10.14**	**96.00**	**0.3771**	**87.27**	12.89	**95.82**	**0.6391**

Table 1 shows a clear gap in speaker-attribution accuracy between the evaluated closed-source and open-source baselines. Averaging equally over the two models in each category, the closed-source baselines lead by 22.64 percentage points on Chinese-language dramas and 36.17 points on English-language series. AutoCAST nevertheless surpasses the strongest closed-source baseline, reaching 91.70% on Chinese-language dramas and 87.27% on English-language series—8.64 percentage points above Doubao-Seed-2.0-lite and 3.53 points above Gemini 3.5 Flash, respectively. AutoCAST also achieves the highest single-speaker coverage and lowest boundary error in both language groups, indicating that improved attribution is accompanied by better temporal segmentation. Despite task-specific training exclusively on Chinese-language dramas, its performance on the evaluated English-language series further supports generalization across languages and TV series.

4.3 Ablation Studies

4.3.1 Transcription Accuracy and Speaker Coverage

To assess the quality of utterance units for downstream attribution, we compare Stage 1 with VibeVoice (Peng et al., 2026), Qwen3-Omni (Xu et al., 2025), and MOSS (AI et al., 2026).

Table 2 shows that, compared with its base model Qwen3-Omni, AutoCAST (Stage 1) reduces AAS from 4.6438 to 0.3771 seconds and increases SSGC from 91.03% to 96.00% (+4.97 pp), improving both boundary precision and single-speaker coverage. Accurate boundaries provide an essential foundation for subsequent voiceprint extraction and reference bank construction, while single-speaker coverage directly affects identity attribution: because each predicted segment receives only one identity label, merging speech from different speakers into a single segment can cause attribution errors. Although MOSS achieves slightly better CER and AAS, AutoCAST combines low boundary error with the highest SSGC (96.00%). To further assess the downstream impact of Stage 1 automatic segmentation, we compare predicted and GT utterance boundaries using GT-labeled reference banks (Table 2(b)). Final attribution accuracy is 92.27% and 92.92%, respectively, a gap of only 0.65 percentage points. This indicates that, under these experimental conditions, downstream attribution using Stage 1 automatic segmentation approaches the performance obtained with GT boundaries.

Table 2: **Transcription quality and downstream segmentation impact.** All metrics are percentages except AAS (seconds); (b) uses GT-labeled reference banks.

(a) Transcription quality

Method	CER↓	SSGC↑	AAS (s)↓
VibeVoice	25.33	67.97	0.8587
MOSS	**9.66**	95.73	**0.3480**
Qwen3-Omni	13.82	91.03	4.6438
AutoCAST (Stage 1)	10.14	**96.00**	0.3771

(b) Downstream segmentation impact

Seg.	Reference bank	Attribution acc.↑
Pred	Manual Labels	92.27
GT	Manual Labels	92.92

4.3.2 Reference Bank Quality and Its Impact

To evaluate Stage 2 reference bank construction, we ablate identity initialization and prediction–face agreement filtering, measuring both seed quality and downstream attribution accuracy. Full Construction uses both components; w/o Face Agreement removes agreement filtering; w/o Init constructs seeds directly from visual identities. All automatic variants use the same predicted segmentation, speaker encoder, propagation backend, and downstream AutoCAST (VL) checkpoint, so their differences reflect the reference bank construction procedure.

Table 3: **Effect of reference bank construction on seed quality and speaker attribution accuracy (%).**

Variant	Seg.	SeedQual↑	Attribution acc. Prop.↑	Attribution acc. Final↑
Full Construction	Pred	**90.89**	**88.63**	**91.70**
w/o Face Agreement	Pred	84.67	84.68	89.75
w/o Init	Pred	69.03	81.88	88.21
Manual Labels	Pred	100.00	88.32	92.27

Table 3 shows that Full Construction achieves the best automatic results, with 90.89% SeedQual. Compared with w/o Init, it improves SeedQual by 21.86 percentage points, propagation accuracy by 6.75 points, and final attribution accuracy by 3.49 points. Compared with w/o Face Agreement, the corresponding gains are 6.22, 3.95, and 1.95 points. These results show that both identity initialization and agreement filtering improve seed quality, and that the resulting reference banks benefit downstream attribution. The Manual Labels configuration provides a comparison using a manually labeled reference bank. Under the same predicted segmentation, final attribution accuracy is 91.70% with the automatically constructed bank and 92.27% with the manually labeled bank, a gap of 0.57

percentage points. These results show that the automatic bank approaches the final attribution performance of the manually labeled bank in this setting. Per-drama results are in Appendix B.2.1.

4.3.3 Multimodal Attribution and Modality Contributions

We evaluate the configurations in Table 4 with GT segmentation and a shared scoring protocol. All attribution models receive GT target text and dialogue context, and all voiceprint-based configurations share a manually labeled reference bank. AutoCAST (Omni) uses two separately fine-tuned Qwen3-Omni models (Xu et al., 2025): one receives audio and images, while the other receives images and reference-derived voiceprint scores. DramaSR-LRM (Li et al., 2026) combines Qwen3-VL-32B with a fine-tuned Qwen3-8B reasoning model (Yang et al., 2025). AutoCAST (VL) uses a single fine-tuned Qwen3-VL-32B model (Bai et al., 2025) to jointly interpret images, voiceprint scores, and dialogue context. Two modality ablations remove images or voiceprint scores while retaining dialogue context; a third ablation removes the multimodal reasoning step and relies solely on acoustic propagation for identity assignment.

Table 4 shows that the reference-assisted AutoCAST (Omni) configuration achieves 90.22% accuracy, compared with 74.42% for direct audio-visual input with the same dialogue context, supporting the performance benefit of the reference-assisted configuration. By combining images, voiceprint scores, and dialogue context, AutoCAST (VL) achieves 92.92% accuracy, surpassing DramaSR-LRM (90.22%) without requiring the additional reasoning model used by DramaSR-LRM. Removing multimodal reasoning, images, or voiceprint scores from AutoCAST (VL) reduces accuracy from 92.92% to 88.49%, 91.49%, and 76.55%, respectively. Together, these results demonstrate the importance of integrating acoustic evidence, visual observations, and dialogue context for speaker attribution.

Table 4: **Speaker attribution accuracy (%) across models and input modalities.**

Method	**Input / Variant**	**Acc.↑**
AutoCAST (Omni)	Audio + images + dialogue context	74.42
AutoCAST (Omni)	Images + voiceprint scores + dialogue context	90.22
DramaSR-LRM (agent)	Images + voiceprint scores + dialogue context	90.22
AutoCAST (VL)	**Images + voiceprint scores + dialogue context**	**92.92**
w/o multimodal reasoning	Acoustic propagation only	88.49
w/o images	Voiceprint scores + dialogue context	91.49
w/o voiceprint scores	Images + dialogue context	76.55

5 Conclusion

We presented AutoCAST, a three-stage framework for automated named-speaker transcription in long-form TV dramas. Timestamp-aware transcription jointly predicts utterance boundaries and text, providing speech segments and dialogue context for subsequent attribution. Agreement-guided reference bank construction combines multimodal identity initialization with prediction–face agreement to select character-specific voice references. Reference-guided multimodal attribution then uses acoustic seed selection and label propagation to obtain character-level voiceprint scores, which are interpreted jointly with visual and dialogue context to determine speaker identities. Together, these stages produce timestamped, character-attributed transcripts from raw audiovisual streams, a cast roster, and character portraits, without manual transcripts or manually verified voice references at inference. On seven held-out Chinese-language dramas and three held-out English-language series from DramaSR-532K, AutoCAST achieves mean speaker-attribution accuracies of 91.70% and 87.27%, exceeding the strongest evaluated closed-source baseline in each language group by 8.64 and 3.53 percentage points, respectively. Ablations support the effectiveness of automatic segmentation for downstream attribution, the contributions of identity initialization and agreement filtering to reference quality, and the complementary roles of acoustic and visual evidence in final attribution. The English-language results further support generalization across languages and TV series following task-specific training on Chinese-language dramas only.

AI USE STATEMENT

We used Gemini to construct reasoning training data for identity initialization and final speaker attribution. The generated examples were filtered against the human-annotated dataset, and the target identity answers were taken from its ground-truth annotations.

We also used large language models to assist with manuscript polishing and formatting, primarily to improve English-language writing and expression. The substantive content of the manuscript was written by the authors.

REFERENCES

MOSI AI, Donghua Yu, Zhengyuan Lin, Hanfu Chen, Chen Yang, Yiyang Zhang, Jingqi Chen, Ke Chen, Liwei Fan, Yi Jiang, et al. Moss transcribe diarize technical report. *arXiv preprint arXiv:2601.01554*, 2026.

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *Interspeech 2023*, pp. 4489–4493, 2023. doi: 10.21437/Interspeech.2023-78.

Hervé Bredin. pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *24th Interspeech Conference (INTERSPEECH 2023)*, pp. 1983–1987. ISCA, 2023.

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

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 16845–16889, 2025.

Xinlong Chen, Yue Ding, Weihong Lin, Jingyun Hua, Linli Yao, Yang Shi, Bozhou Li, Qiang Liu, Yuanxing Zhang, Pengfei Wan, et al. Avocado: An audiovisual video captioner driven by temporal orchestration. In *International Conference on Learning Representations*, volume 2026, pp. 24338–24362, 2026.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Sanjoy Chowdhury, Karren D Yang, Xudong Liu, Fartash Faghri, Pavan Kumar Anasosalu Vasu, Oncel Tuzel, Dinesh Manocha, Chun-Liang Li, and Raviteja Vemulapalli. Amuse: Audio-visual benchmark and alignment framework for agentic multi-speaker understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22998–23009, 2026.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. End-to-end neural speaker diarization with self-attention. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 296–303. IEEE, 2019.

Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Paola Garcia. Encoder-decoder based attractors for end-to-end neural diarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1493–1507, 2022.

Mingyue Huo, Yiwen Shao, and Yuheng Zhang. Tagspeech: End-to-end multi-speaker asr and diarization with fine-grained temporal grounding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41847–41862, 2026.

Koray Kavukcuoglu, Jeff Dean, Oriol Vinyals, and Noam Shazeer. Gemini 3.5: Frontier intelligence with action. Google, May 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/. May 19, 2026.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1369–1379, 2018.

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.

Yuxuan Li, Lingxi Xie, Xinyue Huo, Jihao Qiu, Jiacheng Shao, Pengfei Chen, Jiannan Ge, Kaiwen Duan, and Qi Tian. Reasoning LLM improves speaker recognition in long-form TV dramas. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=h3TLVeukMA.

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. In *International Conference on Learning Representations*, volume 2026, pp. 132277–132304, 2026.

Ivan Medennikov, Taejin Park, Weiqing Wang, He Huang, Kunal Dhawan, Jinhan Wang, Jagadeesh Balam, and Boris Ginsburg. Streaming Sortformer: Speaker Cache-Based Online Speaker Diarization with Arrival-Time Ordering. In *Interspeech 2025*, pp. 5238–5242, 2025. doi: 10.21437/Interspeech.2025-2244.

Arsha Nagrani and Andrew Zisserman. From benedict cumberbatch to sherlock holmes: Character identification in tv series without a script. In *British Machine Vision Conference*, 2017.

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, June 2026.

Zhiliang Peng, Jianwei Yu, Yaoyao Chang, Zilong Wang, Li Dong, Yingbo Hao, Yujie Tu, Chenyu Yang, Wenhui Wang, Songchen Xu, et al. Vibevoice-asr technical report. *arXiv preprint arXiv:2601.18184*, 2026.

Haran Raajesh, Naveen Reddy Desanur, Zeeshan Khan, and Makarand Tapaswi. Micap: A unified model for identity-aware movie descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14011–14021, June 2024.

Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4492–4496. IEEE, 2020.

Mohan Shi, Xiong Xiao, Ruchao Fan, Shaoshi Ling, and Jinyu Li. Train short, infer long: Speech-llm enables zero-shot streamable joint asr and diarization on long audio. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17442–17446. IEEE, 2026.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5329–5333. IEEE, 2018.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-SALMONN: Speech-enhanced audio-visual large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47198–47217. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/sun24l.html.

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. *arXiv preprint arXiv:2502.11775*, 2025.

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.

Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3927–3935, 2021.

Thomas Thebaud, Yuzhe Wang, Laureano Moro-Velázquez, Jesús Villalba-Lopez, and Najim Dehak. Speaker Verification with Speech-Aware LLMs: Evaluation and Augmentation. In *Interspeech 2026*, pp. 4576–4581, 2026. doi: 10.21437/Interspeech.2026-2670.

Quan Wang, Yiling Huang, Guanlong Zhao, Evan Clark, Wei Xia, and Hank Liao. DiarizationLM: Speaker Diarization Post-Processing with Large Language Models. In *Interspeech 2024*, pp. 3754–3758, 2024a. doi: 10.21437/Interspeech.2024-209.

Xizi Wang, Feng Cheng, and Gedas Bertasius. Loconet: Long-short context network for active speaker detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18462–18472. IEEE, 2024b.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 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.

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.

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. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Evaluation Protocols and Metrics

This appendix defines the prediction-to-GT alignment procedure, evaluation metrics, and aggregation rules used for all systems. Unless otherwise specified in this paper, all remaining evaluation criteria follow the DramaSR-532K evaluation protocol (Li et al., 2026).

A.1 Prediction-to-GT Alignment

For predicted utterances, all stages use the same many-to-many matching procedure based on character alignment and temporal gating. Experiments that directly provide GT targets associate outputs by target identifier. Within each evaluation unit, whitespace and the configured Chinese and English punctuation are removed, while letter case is preserved. GT texts and predicted texts are concatenated separately in their original order. We use `SequenceMatcher(None, gt_concat, pred_concat, autojunk=False)` to construct a monotonic character mapping and project each GT utterance's character span onto the predicted text to obtain its initial correspondence.

Up to three predicted segments on either side of the primary prediction are searched. Adjacent predictions are jointly associated with a GT utterance when their concatenated text has similarity of at least 0.55 to that utterance and improves upon the single-segment similarity by at least 0.08.

Before temporal matching, optional unit-level drift correction uses pairs with text similarity of at least 0.6 to calculate the median start-time difference $\Delta = \mathrm{median}(s_g - \hat{s}_p)$. Correction is applied only when there are at least three such pairs, $|\Delta| \geq 1$ second, and the median absolute deviation of the time differences is at most 1 second. In that case, Δ is subtracted from all GT start and end times in the unit for matching; otherwise, the original times are retained. Matched pairs must have overlapping time intervals:

$$\mathrm{IoU}([\widetilde{s}_g, \widetilde{e}_g], [\hat{s}_p, \hat{e}_p]) > 0. \tag{11}$$

Hats denote predicted start and end times, while tildes denote GT start and end times after optional drift correction.

A.2 Single-Speaker GT Coverage (SSGC)

SSGC is defined in Equation 9 in Section 4.1. It measures the fraction of GT utterances covered by at least one matched predicted segment that does not mix different speakers.

Let $\mathcal{U}_{\mathrm{GT}}$ denote all evaluated GT utterances and $\mathcal{G}_p$ the GT utterances matched to prediction p. Define $\mathcal{P}_{\mathrm{ss}}$ as the matched predictions whose associated GT utterances share one speaker identity. Then

$$\mathrm{SSGC} = \frac{|\mathcal{C}_{\mathrm{ss}}|}{|\mathcal{U}_{\mathrm{GT}}|}, \qquad \mathcal{C}_{\mathrm{ss}} = \bigcup_{p \in \mathcal{P}_{\mathrm{ss}}} \mathcal{G}_p. \tag{12}$$

The union counts each covered GT utterance exactly once. A single-speaker predicted segment associated with three GT utterances contributes all three, while multiple qualifying predictions associated with the same GT utterance do not increase its contribution. A GT utterance is counted if at least one associated prediction is single-speaker, even if another associated prediction mixes speakers. Unmatched GT utterances and those covered only by mixed-speaker predictions contribute zero to the numerator but remain in the denominator. Units with GT but no predictions retain all their GT utterances in the denominator. Speaker mixing is determined from GT identities, independently of whether the predicted character name is correct. Higher SSGC indicates greater GT coverage by single-speaker predicted segments.

A.3 Mean Absolute Start- and End-Boundary Error (AAS)

Let $\mathcal{P}_m$ denote the set of predicted segments matched to at least one GT utterance. For $p \in \mathcal{P}_m$, let $s_p^\star$ and $e_p^\star$ be the earliest GT start time and latest GT end time among its associated utterances, both using the original GT timestamps. Then

$$\mathrm{AAS} = \frac{1}{2|\mathcal{P}_m|} \sum_{p \in \mathcal{P}_m} \left(|\hat{s}_p - s_p^\star| + |\hat{e}_p - e_p^\star| \right). \tag{13}$$

AAS averages the absolute errors of the start and end boundaries. Unmatched predictions and GT utterances are excluded. Temporal drift correction is used only to establish matching relationships; it does not alter the original GT timestamps used to calculate AAS. AAS is reported in seconds, while SSGC is converted to a percentage.

A.4 Character-Macro Seed Quality

SeedQual measures the accuracy of initial seed identity labels. We first calculate seed precision for each character and then average over all evaluation characters in the drama, assigning zero to characters without seeds. Only seeds retained before propagation are evaluated; records added during propagation are excluded. All methods use the same evaluation character set.

Following Equation 10 in Section 4.1, let C be the complete evaluation character set for a drama. For construction method m, let $n_{m,c}$ denote the number of seeds assigned to character c, of which $k_{m,c}$ have correct identity labels. We define

$$\mathrm{SeedQual}_m = \frac{1}{|C|} \sum_{c \in C} P_{m,c} \times 100\%, \qquad P_{m,c} = \frac{k_{m,c}}{\max(1, n_{m,c})}. \tag{14}$$

When a character has no seeds, $n_{m,c} = k_{m,c} = 0$, so its precision is zero; it remains in the averaging denominator $|C|$.

A.5 Speaker Attribution Accuracy

Accuracy is calculated over a fixed target set of GT utterances U_d for drama d, with $N_d = |U_d|$. All methods are scored on this set, and each GT utterance is counted once.

A GT utterance may be associated with multiple predictions. It is correct only if it matches at least one prediction and all associated predictions have correct identities. An unmatched GT utterance, or one with any associated prediction whose identity is incorrect, empty, or invalid, is scored as incorrect. If a prediction is associated with multiple GT utterances, it is judged separately against each GT identity. Experiments that directly provide GT targets establish correspondences by target identifier and apply the same scoring rule.

Let $c_{d,i}^{(t)}$ be the correctness indicator for GT utterance i at stage t, equal to 1 when correct and 0 otherwise. Accuracy is the percentage of GT utterances scored as correct:

$$A_d^{(t)} = \frac{100}{N_d} \sum_{i \in U_d} c_{d,i}^{(t)}. \tag{15}$$

GT utterances without predictions remain in the denominator, and extra predictions do not increase it. Evaluation units without GT are excluded.

A.6 Cross-Drama Aggregation

Metrics are averaged equally across series within each language group. For language group $\ell \in \{\mathrm{Zh}, \mathrm{En}\}$, let D_ℓ be the number of evaluated series and $M_{\ell,d}$ the metric for series d. Then

$$\overline{M}_\ell = \frac{1}{D_\ell} \sum_{d=1}^{D_\ell} M_{\ell,d}, \qquad D_{\mathrm{Zh}} = 7, \quad D_{\mathrm{En}} = 3. \tag{16}$$

For the per-drama tables, cross-drama means are calculated from the reported per-drama values and rounded to two decimal places. AAS is displayed to four decimal places in the main comparison and transcription ablation tables, and to two decimal places in the English per-series table. Differences between percentages are reported in percentage points.

B Detailed Experimental Settings and Results

B.1 Ablation and Comparison Settings

The complete pipeline uses Stage 1 predicted utterance boundaries and transcripts. The controlled comparisons below specify when GT segmentation, GT text, or manually labeled reference banks are used. Scoring alignment follows Appendix A.1.

Transcription and segmentation comparisons. Section 4.3.1 evaluates Stage 1 transcription quality and its downstream segmentation impact. Table 2(a) compares CER, SSGC, and AAS across transcription models. Table 2(b) compares predicted (Pred) and GT segmentation using manually labeled reference banks (Manual Labels). The corresponding per-drama final-attribution results appear in the Manual Labels columns of Table 7.

Reference bank construction ablations. The Stage 2 ablations in Section 4.3.2 and Table 3 fix predicted segmentation, the speaker encoder, the propagation backend, and the downstream AutoCAST (VL) checkpoint. Full Construction uses the complete procedure in Section 3.2; w/o Face Agreement removes prediction–face agreement filtering; w/o Init removes audio-visual identity initialization and assigns seeds using visual identities. Manual Labels provides a manually labeled reference-bank comparison under the same predicted segmentation. The per-drama tables additionally report Full Construction and Manual Labels with GT segmentation: the former provides a supplementary segmentation comparison for the automatically constructed bank, and the latter corresponds to the GT condition in Table 2(b).

Attribution model and modality comparisons. Section 4.3.3 and Table 4 compare attribution models using GT segmentation, GT target text, dialogue context, and a shared scoring protocol. The AutoCAST attribution models use the same training data, identity supervision, and training configuration. AutoCAST (Omni) comprises two separately fine-tuned Qwen3-Omni models: one receives target audio and images, and the other receives images and character-specific voiceprint scores. AutoCAST (VL) uses a fine-tuned Qwen3-VL-32B model with images and voiceprint scores. Both input formats also include the shared GT target text and dialogue context. DramaSR-LRM combines Qwen3-VL-32B with a fine-tuned Qwen3-8B reasoning model (Li et al., 2026). All voiceprint-based configurations in this comparison share the Manual Labels reference bank with GT segmentation; the direct audio-and-images configuration uses no acoustic reference bank. The w/o images and w/o voiceprint scores variants each remove the named input from AutoCAST (VL) while retaining dialogue context. The w/o multimodal reasoning variant assigns identities using acoustic propagation alone, with the same manually labeled reference bank.

B.2 Per-Drama Results

Tables 5, 6, and 7 expand the reference-bank and segmentation comparisons in Tables 3 and 2(b). Table 8 expands the four model configurations in Table 4; the three modality and reasoning ablations are reported only as aggregate results in the main text. All per-drama results use the scoring rules in Appendix A.

B.2.1 Agreement-Guided Reference Bank Construction Results

Tables 5, 6, and 7 report seed quality, propagation accuracy, and final attribution accuracy, respectively. Table 5 reports four automatic configurations: w/o Init and w/o Face Agreement with Pred segmentation, and Full Construction with Pred or GT segmentation. Tables 6 and 7 add Manual Labels with Pred or GT segmentation, for six configurations in total. Their four Pred configurations correspond to Table 3. All scores are percentages, with equal-weight seven-drama means. SeedQual follows Equation 10, including zero contributions from unseeded roster characters.

Table 5: **Per-drama seed quality (%).** Roster is the GT character count, with its mean rounded to an integer. Duration (h:mm) and utterance counts are dataset statistics; their last-row entries are totals.

				Ablations (Pred)		Full Construction	
Drama	**Duration** (h:mm)	**#uttr.**	**Roster**	**w/o Init**	**w/o Face Agreement**	**Pred**	**GT**
The Long Night（沉默的真相）	10:20	8,263	37	67.18	91.89	94.59	94.59
Qin Empire 2（大秦帝国之纵横）	38:42	29,127	60	77.00	68.33	83.67	80.00
Ode to Joy 1（欢乐颂1）	30:04	35,949	59	54.24	76.27	81.36	79.66
The Knockout（狂飙）	27:02	36,183	75	74.00	89.33	94.67	93.33
Minning Town（山海情）	17:34	23,932	85	69.41	82.35	88.24	90.59
Stand by Me 1（一起同过窗1）	25:18	33,749	30	76.67	93.33	96.67	100.00
Battle of Changsha（战长沙）	23:31	20,680	34	64.71	91.18	97.06	97.06
Mean / total	172:31	187,883	54	69.03	84.67	90.89	90.75

Table 6: **Per-drama propagation accuracy (%).** Results before reference-guided multimodal attribution.

	Ablations (Pred)		Full Construction		Manual Labels	
Drama	**w/o Init**	**w/o Face Agreement**	**Pred**	**GT**	**Pred**	**GT**
The Long Night（沉默的真相）	69.19	86.58	85.69	84.56	87.75	87.77
Qin Empire 2（大秦帝国之纵横）	83.54	74.89	87.07	85.51	83.71	84.28
Ode to Joy 1（欢乐颂1）	86.91	89.00	89.78	88.53	90.53	91.87
The Knockout（狂飙）	79.77	84.12	87.57	88.44	87.60	87.76
Minning Town（山海情）	81.41	79.21	88.16	88.84	87.32	88.43
Stand by Me 1（一起同过窗1）	88.61	93.03	93.36	93.71	93.40	93.95
Battle of Changsha（战长沙）	83.74	85.92	88.79	84.31	87.95	85.37
Seven-drama mean	81.88	84.68	88.63	87.70	88.32	88.49

Table 7 reports final multimodal attribution accuracy for six reference-bank configurations. w/o Init and w/o Face Agreement use predicted segmentation. All configurations follow the definitions in Tables 5 and 6.

Table 7: **Per-drama final-attribution accuracy (%).** The multimodal attribution model is fixed across reference-bank configurations.

	Ablations (Pred)		Full Construction		Manual Labels	
Drama	**w/o Init**	**w/o Face Agreement**	**Pred**	**GT**	**Pred**	**GT**
The Long Night（沉默的真相）	83.07	90.58	91.21	90.07	90.11	92.54
Qin Empire 2（大秦帝国之纵横）	88.84	80.85	90.76	90.49	91.70	92.56
Ode to Joy 1（欢乐颂1）	90.69	92.48	91.89	92.25	95.06	95.14
The Knockout（狂飙）	87.92	92.01	92.45	92.18	93.09	93.66
Minning Town（山海情）	85.91	86.66	90.97	90.08	90.10	90.61
Stand by Me 1（一起同过窗1）	92.00	94.77	95.10	94.94	95.03	95.64
Battle of Changsha（战长沙）	89.03	90.93	89.54	91.37	90.79	90.29
Seven-drama mean	88.21	89.75	91.70	91.63	92.27	92.92

B.2.2 Reference-Guided Multimodal Attribution Results

All attribution models are evaluated on the shared targets from seven held-out dramas, with identical GT target transcription texts and dialogue context. Reference-conditioned attribution models share the Manual Labels GT bank $\mathcal{B}^{\mathrm{GT}}$; the AutoCAST (Omni) audio configuration receives no acoustic references.

Table 8: **Per-drama reference-guided multimodal attribution accuracy (%).** Bold denotes the best result in each row.

Drama	**AutoCAST (Omni) Audio**	**AutoCAST (Omni) Scores**	**AutoCAST (VL) Scores**	**DramaSR-LRM**
The Long Night（沉默的真相）	74.27	88.97	**92.54**	90.15
Qin Empire 2（大秦帝国之纵横）	62.33	89.27	**92.56**	88.24
Ode to Joy 1（欢乐颂1）	83.68	93.06	**95.14**	93.34
The Knockout（狂飙）	78.56	91.31	**93.66**	89.87
Minning Town（山海情）	76.89	88.46	**90.61**	88.44
Stand by Me 1（一起同过窗1）	76.98	93.74	**95.64**	93.63
Battle of Changsha（战长沙）	68.24	86.76	**90.29**	87.84
Seven-drama mean	74.42	90.22	**92.92**	90.22

AutoCAST (Omni) (Audio) uses audio and images; AutoCAST (Omni) (Scores) and AutoCAST (VL) use images and structured voiceprint scores. All attribution models also receive GT target text and dialogue context. DramaSR-LRM uses Qwen3-VL-32B + Qwen3-8B (Li et al., 2026).

Qualitative examples. Figure 3 illustrates three complementary decision patterns. In the first example, the character with the highest voiceprint similarity is also consistent with the visible speaker and her on-screen behavior, so the acoustic and visual evidence reinforce the same attribution. In the second example, a clear voiceprint-score margin disambiguates the speaker among the visible candidates, while the scene context supports the inferred interaction. In the third example, several candidates receive near-tied voiceprint scores, making acoustic evidence alone insufficient; the visible participant relations and the semantics of the utterance then support selecting Guan. Together, these cases show that acoustic cues, visual observations, and dialogue context provide complementary evidence for speaker attribution.



Figure 3: **Qualitative examples of multimodal speaker attribution. Left:** Consistent face and voice evidence confirms the speaker. **Middle:** A clear voiceprint-score margin resolves identity ambiguity. **Right:** When voiceprint scores are close, visible participants and dialogue semantics enable context-aware correction.

B.3 Cross-Language Generalization

To assess performance beyond Chinese-language dramas, we evaluate the complete pipeline on the three English-language series introduced in Section 4.1. Table 9 reports WER, AAS, SSGC, SeedQual, and attribution accuracy after Init, acoustic propagation, and final multimodal attribution.

Table 9: **Results on English-language TV series.** Metrics are percentages except AAS (s). Duration (h:mm) and utterance counts are dataset statistics. Panel (a) reports dataset statistics and totals; panel (b) reports performance with equal-weight means across series.

(a) Dataset statistics

Series	Seasons	Duration (h:mm)	#uttr.
Friends	10	89:25	95,564
Lost	5	87:33	66,688
Downton Abbey	6	51:08	56,583
Total	21	228:06	218,835

(b) Transcription and speaker attribution

	Transcription			Initialization / seeds		Attribution acc.	
Series	**WER↓**	**AAS↓**	**SSGC↑**	**Init.↑**	**SeedQual↑**	**Prop.↑**	**Final↑**
Friends	15.09	0.55	97.08	81.62	86.49	83.89	86.15
Lost	15.84	0.89	91.68	78.04	85.16	80.98	83.92
Downton Abbey	7.73	0.48	98.71	80.83	98.02	91.34	91.75
Mean	**12.89**	**0.64**	**95.82**	**80.16**	**89.89**	**85.40**	**87.27**

Table 9 shows that mean attribution accuracy rises from 80.16% after Init to 85.40% after propagation and 87.27% after final multimodal attribution, a total gain of 7.11 points. Final multimodal attribution improves over acoustic propagation on all three series. These results demonstrate the value of reference banks on English-language content.

C METHODS AND MODEL TRAINING

C.1 FRAMEWORK AND CLIP PROCESSING

Face-annotation preprocessing. The available inputs are raw video and audio (V, A), a cast roster $\mathcal{C} = \{c_k\}_{k=1}^{K}$, and corresponding reference portrait sets $\mathcal{P} = \{F_k\}_{k=1}^{K}$, where F_k contains portraits of character c_k. Following DramaSR-LRM (Li et al., 2026), we use V, $\mathcal{C}$, and $\mathcal{P}$ to obtain the character-annotated video $\widetilde{V}$. The annotations associate detected face boxes with character names and face-recognition confidence scores. AutoCAST then receives $(\widetilde{V}, A, \mathcal{C})$ and uses the preprocessed video as visual identity evidence in the three-stage framework.

Input windows and target units. For a drama of duration D, inference starts at $o_1 = 0$. The b-th transcription input window spans at most 60 seconds:

$$\mathcal{I}_b = [o_b, e_b)\,, \qquad e_b = \min(o_b + 60, D). \tag{17}$$

Let M_b be the number of utterances predicted in this window before retention, and let $\widehat{e}_{b,j}$ be the clip-local end time of its j-th prediction. If $e_b < D$ and $M_b \geq 2$, we retain the first $M_b - 1$ predictions, discard the last, and set

$$o_{b+1} = o_b + \widehat{e}_{b,M_b-1}. \tag{18}$$

Thus, the next window starts at the end of the penultimate predicted utterance, allowing the discarded final utterance's audio to be processed again. We use *input window* (or *clip*) for the audio-visual interval $\mathcal{I}_b$, and *unit* for its retained target sequence. The retained utterances define unit b: the discarded last prediction is not part of this unit. For a non-final window, the unit interval is $[o_b, o_{b+1})$; the final unit ends at D. Thus, unit intervals do not overlap, and each retained utterance belongs to exactly one unit. The transcription input window $\mathcal{I}_b$ may extend beyond the unit boundary to provide the look-ahead needed to choose that boundary. Discarded predictions are not included in the merged transcript.

If a non-final window has no predictions, it contributes no records and the next window starts at e_b. If it has exactly one prediction, the implementation raises an error because no penultimate endpoint exists. Each new offset must strictly advance. When $e_b = D$, all predictions are retained and processing ends, including when the final window is empty or contains only one utterance.

Stages 2 and 3 reuse these window intervals and the retained target records assigned to each unit; units without retained targets are skipped for attribution. Frames and audio are selected using the actual window offsets. Clip-relative timestamps of retained records are converted to drama-level timestamps as

$$\tau_{b,i}^{\text{global}} = o_b + \tau_{b,i}^{\text{local}}. \tag{19}$$

The shared bank pools evidence from the entire target drama, including future clips. Utterance-level initial identities are predicted by the Init model. This is an offline, transductive protocol.

C.2 TIMESTAMP-AWARE TRANSCRIPTION

Segmentation instruction. Stage 1 uses the transcription instruction P_{seg} and the timestamp-aware audio sequence A_b^{ts} to generate a JSON array containing start times, end times, and text for clip b. As in Section 3.1, integer-second text markers t_j are placed before raw audio segments $A_{b,j}$ for the local intervals $[j, j+1)$ during input construction, before audio encoding.

Figure 4 presents the transcription prompt.

Stage 1: Timestamp-aware transcription

Task

You are an automatic speech recognition (ASR) assistant for film and television scenes. Transcribe all clearly intelligible spoken dialogue and segment it using natural punctuation.

Input

- Audio is provided as successive audio segments with interleaved integer-second text markers.
- Each time marker indicates the local time in seconds from the beginning of the clip and precedes the audio segment for that second.
- The markers provide temporal references only. Listen to the audio as one continuous recording; do not infer pauses or breaks from the markers.
- No subtitles, images, or reference answers are provided.

```
Time {t}
<audio segment for [t, t+1)>
... (repeat time markers and
audio segments in
chronological order)
```

Transcription and segmentation

- Follow speaking order and segment by natural punctuation, including periods, question marks, exclamation marks, and commas.
- A sentence may span several audio segments. Keep it together; do not split a sentence at time markers.
- Use `...` for unintelligible words. Skip an entire utterance if it cannot be heard.
- Exclude background music, environmental sounds, laughter, and other content without clear dialogue.
- Do not invent content that was not heard.

Timestamps

- Align each start and end time as closely as possible with the actual onset and offset of speech, using two decimals.
- Use the interleaved time markers as references to locate utterance boundaries. Do not guess times or round them arbitrarily.
- Utterance intervals must be chronologically ordered and non-overlapping. Natural pauses between utterances are allowed; timestamps must not reverse.

Output format

Return only a JSON array, without Markdown fences or extra text. Each object must contain `start`, `end`, and `text`:

```
[
  {"start": 0.50, "end": 2.06,
   "text": "Ningxia Hui
   Autonomous Region"},
  {"start": 2.34, "end": 3.14,
  "text": "for Xihaigu"}
]
```

Figure 4: Timestamp-aware transcription prompt (Stage 1).

Structured supervision. Each annotated utterance is represented by a record:

$$r_i = \{\texttt{start} : \tau_i^{\text{start}},\ \texttt{end} : \tau_i^{\text{end}},\ \texttt{text} : y_i\}. \tag{20}$$

The training target is the JSON serialization of the temporally ordered record list:

$$Y = \text{JSON}([r_1, r_2, \ldots, r_N]), \tag{21}$$

Keys and transcript values are JSON strings, while timestamps are numeric values with two decimal places. Each JSON object corresponds to one annotated utterance. The target contains neither character names nor anonymous speaker indices. Training and inference use the same JSON-array format.

If $Y = (v_1, \ldots, v_M)$ is the tokenized target sequence, the segmentation model is trained with

$$\mathcal{L}_{\text{seg}} = -\sum_{m=1}^{M} \log p_{\theta_{\text{seg}}}\left(v_m \mid v_{<m}, P_{\text{seg}}, A^{\text{ts}}\right). \tag{22}$$

No separate timestamp-regression or speaker-classification loss is introduced. At inference, the prompt requests a JSON array. We parse its `start`, `end`, and `text` fields and use the time ranges to extract corresponding audio and visual segments for subsequent identity analysis.

C.3 Audio-Visual Identity Initialization (Init) Model Training

Identity-initialization instruction. The Init model combines video, audio, and utterance records to generate an initial character identity for each target utterance. Its inputs include the attribution instruction P_{attr}, video frames $\widetilde{V}$ annotated with character names and face boxes, the audio sequence A_b^{ts}, and numbered target utterance records. The audio and frames come from the input window associated with the target unit. Up to eight preceding and eight following utterances provide textual context and may lie outside that window. The input presents all timestamped video frames first, followed by the audio sequence and numbered utterance records. Integer-second text markers are interleaved with raw audio segments before audio encoding, using the same input construction as Stage 1.

The instruction requests identity assignments only for numbered target utterances, with contextual dialogue provided as reference. The output is a JSON object containing `answer` and `reasons`, which provide character names and reasoning rationales, respectively, keyed by the same utterance identifiers.

Figure 5 presents the identity-initialization prompt and its input template; placeholders represent instance-specific content.

Stage 2: Audio-Visual Identity Initialization (Init)

Task

You are a speaker-identification assistant for film and television scenes. Use the context dialogue, audio, and sampled frames provided by the user to identify the most likely speaker for each target utterance, and give detailed supporting reasoning.

Input and evidence

The user provides face-annotated video frames followed by the corresponding audio sequence. Use the time markers, images, lip movements, voice characteristics, and context dialogue to identify character names.

Video clip information

- **Clip ID**: Group {unit_idx} (Episode {episode_idx})
- **Clip time range**: {start_time} ~ {end_time}

Reference video frames

The following {num_frames} frames are uniformly sampled from the clip. Each frame is annotated with detected face information in the format `character name, face recognition confidence`.

```
Time {frame_idx}
<image>
... (repeat frame labels and
images in chronological order)
```

Audio

After all reference frames, the corresponding audio is provided as a sequence of successive audio segments with interleaved integer-second text markers. Each marker gives the local time in seconds from the beginning of the clip and precedes the corresponding audio segment. Use these markers to align audio with the reference frames and target utterances; do not treat them as pauses or utterance boundaries.

```
Time {t}
<audio segment for [t, t+1)>
... (repeat time markers and
audio segments in
chronological order)
```

Subtitle list

The subtitles below are in chronological order and include up to 8 context subtitles before and after the target subtitle sequence.

- Subtitles labeled [C-N] or [C+N] are **context subtitles** (reference only; no speaker identification is required).
- Subtitles labeled [1], [2], ... are **target subtitles** (you must identify their speakers). Use each target number as the key in `answer` and `reasons`.

Context subtitles (before the targets)

```
[{label}] Time:
{start_time}-{end_time} |
Text: "{text}"
... (repeat if preceding
context exists; otherwise omit
this part)
```

Target subtitles

```
[{number}] Time:
{start_time}-{end_time} |
Text: "{text}"
... (repeat for each target
subtitle)
```

Context subtitles (after the targets)

```
[{label}] Time:
{start_time}-{end_time} |
Text: "{text}"
... (repeat if following
context exists; otherwise omit
this part)
```

Output requirements

The output must be strict JSON containing only `answer` and `reasons`.

- `answer` is an object whose keys are target-number strings (`"1"`, `"2"`, ...) and whose values are character-name strings.
- `reasons` must have exactly the same keys as `answer` and should cover voice characteristics, lip-movement/audio alignment, faces in sampled frames, contextual continuity, and excluded alternatives wherever possible.

```
{"answer": {...}, "reasons":
{...}}
```

Figure 5: Audio-visual identity-initialization prompt and input template (Stage 2).

Structured supervision. Training uses GT utterance boundaries and transcripts, denoted by $\mathcal{U}^{\mathrm{GT}}$, augmented with up to eight preceding and eight following context utterances as $\mathcal{U}^{\mathrm{GT,ctx}}$. Supervision covers only the target utterances. During training-data construction, Gemini receives the training question and its GT identity answers and generates answer-conditioned rationales. Identity labels come from GT annotations, while the rationales are generated by Gemini.

The training target is a JSON object containing identity answers in `answer` and rationales in `reasons`, keyed by the same target-utterance identifiers. Training and inference use the same output format. Let $W = (w_1, \ldots, w_L)$ be the target sequence obtained by serializing and tokenizing this JSON object. The model is trained with

$$\mathcal{L}_{\mathrm{attr}} = -\sum_{l=1}^{L} \log p_{\theta_{\mathrm{attr}}} \left(w_l \mid w_{<l}, P_{\mathrm{attr}}, \widetilde{V}, A^{\mathrm{ts}}, \mathcal{U}^{\mathrm{GT,ctx}} \right). \tag{23}$$

The loss covers all output tokens in both the identity answers and rationales. The shared training configuration is specified in Section 4.1. At inference, the model receives Stage 1 predicted utterance records and generates its own identities and rationales. We parse `answer` and extract initial identity labels by utterance identifier.

C.4 Reference-Guided Multimodal Attribution: Evidence and Training

Report generation and reuse. At inference, the attribution model's acoustic input is the final `rel_info` report produced by propagation, which provides the name–score reports $\mathbf{k}_i$ described in Section 3.3. For each utterance passed to final attribution, it contains between one and twenty candidate roles:

$$\mathcal{R}_i = ((g_{i,q}, s_{i,q}))_{q=1}^{M_i}, \qquad 1 \le M_i \le 20. \tag{24}$$

Here, $g_{i,q}$ is the role at rank q, and $s_{i,q}$ is its aggregated acoustic similarity. The attribution model directly reuses the backend's reported scores without additional rescoring.

Candidates are first restricted to temporally eligible characters with finite acoustic scores. If this set is empty, the default configuration falls back to all roster characters with finite scores, without temporal filtering. Candidates are then ranked by descending score and truncated to at most twenty. Characters without references are excluded from scoring. If no acoustic seeds survive reference selection, processing stops with an error; final attribution also rejects empty candidate lists.

Report serialization. Reports follow the order of utterance records $\mathcal{U}_b$ in clip b. The candidate names and scores are denoted by $\mathcal{K}_b$. Unlisted roles have no supplied acoustic score; their absence may result from temporal gating or rank truncation and does not indicate a poor match. A score field marked `N/A` denotes unavailable evidence. The final attribution model receives $[P_{\mathrm{final}}; \widetilde{V}_b; \mathcal{U}_b^{\mathrm{ctx}}; \mathcal{K}_b]$, without raw audio, reference recordings, or embedding vectors.

The attribution model combines acoustic scores, temporally aligned video frames, and dialogue context to select a candidate name for each target utterance. This restriction is imposed by the prompt and does not imply constrained decoding. The expanded sequence $\mathcal{U}_b^{\mathrm{ctx}}$ includes the target records $\mathcal{U}_b$ together with up to eight preceding and eight following utterances from $\mathcal{U}$. These surrounding utterances provide context only. The requested output is strict JSON containing rationales and identities for the N_b target utterances in $\mathcal{U}_b$, with no predictions requested for the surrounding context.

Reference-guided multimodal attribution instruction. Figures 6 and 7 present the attribution model prompt and its input template. The prompt refers to utterance texts as subtitles; these are Stage 1 predictions in the complete system and GT texts in the designated reference-assisted conditions. Placeholders represent instance-specific content. The numerical similarity values below are prompt-level heuristics, not hard thresholds enforced by the decoder.

Stage 3: Reference-guided multimodal attribution (part 1 of 2)

Task

You are a professional film and television content analysis system performing speaker identification for a television drama clip. Your task is to combine video frames (with facial identity annotations), subtitle text, and audio similarity cues to identify the speaker of each target subtitle below. Choose one speaker from the candidate list associated with each subtitle as its final speaker.

Video clip information

- **Clip ID**: Group {unit_idx} (Episode {episode_idx})
- **Clip time range**: {start_time} ~ {end_time}

Reference video frames

The following {num_frames} frames are uniformly sampled from the clip. Each frame is annotated with detected face information in the format `character name, face recognition confidence`.

```
Time {frame_idx}
<image>
... (repeat frame labels and
images in chronological order)
```

Subtitle list

The subtitles below are in chronological order and include up to 8 context subtitles before and after the target subtitle sequence.

Notes

- Subtitles labeled [C-N] or [C+N] are **context subtitles** (reference only; no speaker identification is required).
- Subtitles labeled [1], [2], ... are **target subtitles** (you must identify their speakers).

Context subtitles (before the targets)

```
[{label}] Time:
{start_time}-{end_time} |
Text: "{text}"
... (repeat if preceding
context exists; otherwise omit
this part)
```

Target subtitles

```
[{number}] Time:
{start_time}-{end_time} |
Text: "{text}"
    Candidate speakers (audio
    similarity): {candidates}
... (repeat for each target
subtitle)
```

Context subtitles (after the targets)

```
[{label}] Time:
{start_time}-{end_time} |
Text: "{text}"
... (repeat if following
context exists; otherwise omit
this part)
```

Important guidance

Understanding audio similarity

- Candidate speakers for each target subtitle are sorted by audio similarity in descending order (scores appear in parentheses).
- **Focus on relative rankings and score gaps between candidates, rather than mechanically judging against fixed values.** Absolute audio similarity values vary substantially across dramas, recording conditions, and voiceprint systems. There is no precise universal threshold applicable to all scenarios.
- Key signal: the **larger the gap** between the leading candidate and the others, the more reliable the audio cue. The **closer the scores**, the harder it is to distinguish candidates using audio alone; give greater weight to visual evidence and dialogue logic in such cases.
- The following values are rough references only (actual values vary by scenario; consider score gaps and visual evidence):
  - A relatively high top score (roughly above 0.6) suggests that the audio cue is relatively reliable.
  - A relatively low top score (roughly below 0.4) suggests limited confidence in the audio cue; rely more on visual evidence and dialogue logic.
  - If all candidates score within the same low range, overall uncertainty is higher; integrate multiple sources of evidence.

Figure 6: Reference-guided multimodal attribution prompt (Stage 3), part 1: input template and audio-similarity guidance.

Stage 3: Reference-guided multimodal attribution (part 2 of 2)

[Important] Temporal correspondence between frames and subtitles

- Each frame is labeled with its timeline position. Each subtitle's start-end interval likewise specifies its corresponding period.
- **Use visual evidence strictly according to time intervals**: if a character appears or speaks only near an early time A, this **does not** establish that they speak a later subtitle at time B, where B is much later than A.
- For each subtitle, prioritize frames **within or near** its time interval. Do not generalize a visual impression from one period to the entire clip.

[Important] The speaker may not appear in the current frame

- A character remaining on screen (for example, continuously in the foreground) does not necessarily mean that they are speaking. The actual speaker may be off-screen.
- **Do not identify a speaker solely because the character appears on screen.**
- Combine audio similarity (the primary reference), lip movements, facial orientation, dialogue logic, and other evidence. Even if a character dominates the image, follow the audio evidence.

Suggestions for integrated analysis

- Consider which characters appear in the reference frames and their positions in the image.
- Use dialogue logic and scene continuity from context subtitles.
- Note that multiple candidate characters may be present in the same scene.

[Important] Use direct evidence from this clip, not plot knowledge

- Your decisions must primarily rely on **audio similarity and visual frames within this clip**.
- Even if you know the plot, characters' fates, or story trajectory, **do not use plot-based inference to override** the speaker indicated by audio similarity or visual evidence. For local clip analysis, these direct cues are much more reliable than plot logic.
- If both audio similarity and visual information clearly indicate a speaker, trust that direct evidence even if background plot knowledge suggests another possibility.

[Important] Trust the audio ranking in the candidate list

- Candidates are sorted by audio similarity in descending order. When the top-ranked candidate is a **known character**, the audio algorithm considers that person's voiceprint the closest match.
- Without strong contrary visual evidence, do not overturn the audio ranking and select a lower-ranked candidate merely because that character is more visually prominent.
- Audio similarity ranking is the primary reference; visual information supplements rather than replaces it.

Output format

Output the analysis and reasoning process, together with the speaker assignment for each target subtitle, strictly in JSON format.

Strict requirements

- Output only JSON, with no Markdown fences or text outside JSON.
- Speaker assignments must correspond one-to-one with the numbered target subtitles.
- Speaker names must exactly match a name in the corresponding subtitle's candidate list (do not rephrase).
- Select exactly one speaker per subtitle; do not omit any number.

Figure 7: Reference-guided multimodal attribution prompt (Stage 3), part 2: remaining evidence guidance and output requirements. Read after Figure 6.

Supervision and inference. Attribution model training uses GT utterance boundaries and transcripts. Let $\mathcal{U}_b^{\mathrm{GT}}$ denote the ordered annotated records in training clip b, and let $\mathcal{K}_b^{\mathrm{GT}}$ denote their score reports derived from GT character banks. We augment these target records with up to eight preceding and eight following GT utterances, denoting the expanded input by $\mathcal{U}_b^{\mathrm{GT,ctx}}$. The surrounding utterances provide context only; supervision covers the target utterances in $\mathcal{U}_b^{\mathrm{GT}}$. Each target record has an annotated GT identity. During training-data construction, Gemini receives the training question and its GT answer and generates an answer-conditioned rationale. The supervised output contains both rationales and identity answers, with answers following utterance order. Let $(w_1, \ldots, w_L)$ be the tokens of this complete output. The final model minimizes

$$\mathcal{L}_{\mathrm{final}} = -\sum_{l=1}^{L} \log p_{\theta_{\mathrm{final}}}(w_l \mid w_{<l}, P_{\mathrm{final}}, \widetilde{V}_b, \mathcal{U}_b^{\mathrm{GT,ctx}}, \mathcal{K}_b^{\mathrm{GT}}). \tag{25}$$

Upstream model checkpoints and feature extractors remain frozen. During training, the attribution model uses the GT records and bank-derived reports above; in the complete system, it uses Stage 1 records and automatically constructed banks. At inference, the attribution model receives neither GT answers nor Gemini-generated rationales. It generates its own rationales and identity answers for the target utterances. Identity answers, rather than rationale text, are used for attribution scoring and paired with the corresponding transcription records. Record timestamps are converted to global timestamps, and records are concatenated in temporal order.