VOCALCAP: A SPEAKER-LEVEL BENCHMARK FOR FINE-GRAINED SPEECH CAPTIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding speech beyond its words requires describing vocal attributes and associating them with the speakers who produced them. Existing audio-language benchmarks provide limited evaluation of this joint requirement in multi-speaker recordings. We introduce VocalCap, a benchmark for speaker-grounded speech captioning that links fine-grained vocal description with speaker coverage, temporal localization, and attribute-speaker binding. VocalCap comprises 1,000 human-annotated single-speaker clips from film and television and 1,000 controlled mixtures containing two to six speakers. Both tracks share a 13-field schema spanning perceived speaker characteristics, prosody, articulation, affect, and speaking style. Our evaluation combines field-wise comparison against human references with deterministic temporal matching, reporting coverage, timeline accuracy, and binding for multi-speaker outputs. We evaluate 14 systems on the single-speaker track and 12 on the multi-speaker track. Qwen3.5-Omni-Plus leads both, achieving a normalized single-speaker score of 72.2 and a multi-speaker score of 69.7 on their respective 0–100 scales. Cross-track rank changes show that strong single-speaker captioning does not ensure reliable multi-speaker performance. Scores for affect and style are lower than those for gender presentation and accent, although these subjective fields also have lower annotation agreement. These findings highlight the need to assess caption quality together with speaker coverage and temporal grounding. We release the annotations, source pointers, construction metadata, and evaluation code.

1 INTRODUCTION

Audio-language models increasingly support natural-language reasoning and description of speech beyond transcription (Tang et al., 2024; CHEN et al., 2025). In recordings with multiple speakers, a description must also preserve which voice supports each attribute. A model may recognize a low pitch and an urgent tone yet assign these properties to the wrong speakers, or describe the dominant voice accurately while omitting another. These cases motivate evaluating descriptive fidelity together with speaker coverage and temporal correspondence.

Speech-captioning datasets provide detailed descriptions of vocal characteristics and speaking styles (Jin et al., 2024; Diwan et al., 2025), including multi-talker settings (Huang et al., 2025a). Audio understanding benchmarks test speaker, temporal, and compositional relations through questions or audio–caption matching (Sakshi et al., 2025; Ghosh et al., 2024), while diarization evaluates speaker activity over time (Bredin et al., 2020; Ryant et al., 2019). However, success on a selected question or relation does not establish that a model can produce a complete set of fine-grained captions with the correct speaker assignments and activity intervals. Evaluating this complete output requires both descriptive references and explicit checks for missing speakers, inaccurate timing, and incorrect attribute assignments.

We introduce *VocalCap*, a benchmark for *speaker-grounded speech captioning*. Its single-speaker track contains 1,000 human-annotated film and television clips, selected to broaden coverage across 12 vocal attributes and supplemented with a free-form summary. Its multi-speaker track contains 1,000 controlled mixtures of two to six speakers drawn from 735 eligible source clips. The shared 13-field schema covers perceived speaker characteristics, prosody, articulation, affect, and style. Controlled mixing retains known source identities and activity intervals while varying overlap,

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.

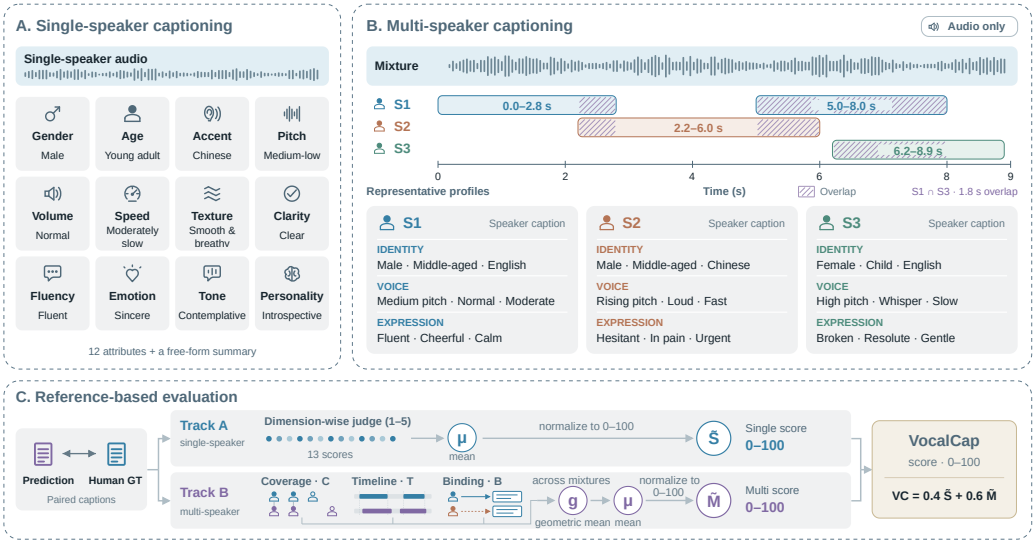

Figure 1: VocalCap evaluates descriptions at two levels: single-speaker field agreement and multi-speaker Coverage, Timeline, and Binding. Multi averages per-mixture geometric scores. The profiles and intervals are schematic examples, not experimental outputs.

speaker similarity, gain, and noise, providing traceable references for evaluating speaker correspondence.

VocalCap evaluates each description in the context of the speaker and intervals to which it is assigned. A text-only judge compares predicted fields with human references, while deterministic temporal matching establishes speaker correspondence independently of caption content. The multi-speaker protocol reports Coverage, Timeline, and Binding, and averages their per-mixture geometric means to obtain the overall score. Binding includes all reference speakers, with zero contribution from unmatched sources. The protocol thus penalizes missing speakers even when captions for recovered speakers are accurate.

We evaluate 14 systems on the single-speaker track and 12 on the multi-speaker track. Qwen3.5-Omni-Plus leads both, with scores of 72.2 and 69.7 on the respective 0–100 scales (Figures 3 and 4). The relative ordering of other systems changes: Qwen3-Omni-Instruct outperforms MiMo-V2.5 on the single-speaker track (68.7 vs. 64.7), but trails it on the multi-speaker track (32.4 vs. 55.4). Single-speaker rankings therefore do not fully characterize performance when descriptions must be assigned across a complete recording. Field-wise results also show lower reference agreement for texture, emotion, tone, and personality than for gender presentation and accent. We interpret these differences alongside the lower human annotation agreement for the former fields (Table S3).

To summarize, our contributions are threefold:

- A benchmark linking fine-grained single-speaker captioning to controlled multi-speaker evaluation through a shared schema and human reference descriptions.
- A protocol that evaluates descriptive agreement jointly with speaker coverage and temporal accuracy, including omitted speakers in caption aggregation.
- An empirical analysis of cross-track rankings, caption reassignment, and reference-structure assistance, interpreted alongside annotation reliability and mixture composition.

2 Related Work

Audio-language understanding. AudioCaps and Clotho established benchmarks for describing acoustic events in natural language (Kim et al., 2019; Drossos et al., 2020). Audio-language models such as SALMONN and Qwen2-Audio extend this interface to transcription, question answering,

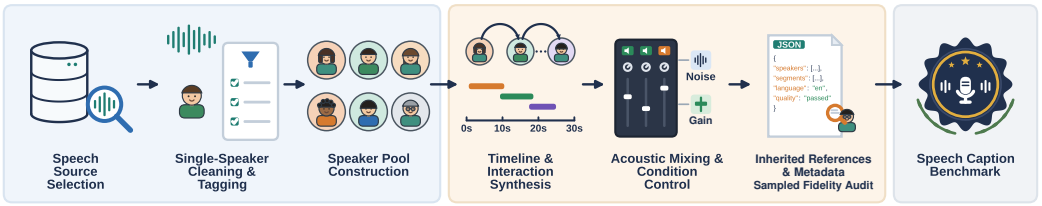

Figure 2: The VocalCap construction pipeline. Human descriptions are retained as references; canonical tags guide source selection. Controlled timelines, gain, and optional noise produce traceable mixtures without artificial reverberation.

and audio understanding (Tang et al., 2024; Chu et al., 2024). AIR-Bench, AudioBench, and MMAU evaluate these capabilities across diverse tasks (Yang et al., 2024; Wang et al., 2025a; Sakshi et al., 2025), including emotion, speaker roles, and temporal reasoning in MMAU. Success on these tasks, however, does not establish whether a model can produce a complete and accurate vocal description for every speaker in a recording.

Speech captioning and descriptive supervision. StyleCap and factor-conditioned captioning generate natural-language descriptions of speaking style (Yamauchi et al., 2024; Ando et al., 2024). LibriTTS-P, SpeechCraft, and ParaSpeechCaps expand the available speech–description pairs for captioning and controllable synthesis (Kawamura et al., 2024; Jin et al., 2024; Diwan et al., 2025); CapSpeech further benchmarks downstream applications of style-captioned text-to-speech (Wang et al., 2026). MOSS-Audio provides a particularly close evaluation protocol: models generate captions along 13 dimensions, and a text-based judge scores each dimension against human-written references (Yang et al., 2026a). VocalCap places such field-wise description within a shared evaluation framework for single-speaker clips and controlled multi-speaker mixtures.

Paralinguistic and perceptual evaluation. CP-Bench evaluates contextual reasoning that combines linguistic content with paralinguistic cues (Wang et al., 2025b). Descriptive speech-quality evaluation uses natural language to characterize perceptual signal quality (CHEN et al., 2025), while SpeechParaling-Bench and ParaS2S assess paralinguistic speech generation and speech-to-speech interaction, respectively (Liu et al., 2026; Yang et al., 2026b). VocalCap focuses on the fidelity of vocal descriptions grounded in the input recording.

Speaker attribution and temporal grounding. SpeechCaps uses multi-talker speaking-style captions for training and evaluates associations between speakers and vocal attributes through generated answers to questions (Huang et al., 2025a). CompA probes temporal composition and attribute binding through audio–caption matching (Ghosh et al., 2024), while diarization evaluates speaker activity over time (Ryant et al., 2019; Fiscus et al., 2006). VocalCap requires a complete structured response containing one item per speaker, all activity intervals, and a caption with 12 vocal attributes and one summary. Its multi-speaker protocol reports Coverage, Timeline, and Binding: temporal matching establishes speaker correspondence before field-wise caption scoring, and unmatched reference speakers receive zero binding credit.

3 The VocalCap Benchmark

3.1 Task Definition

Given an audio clip x, the single-speaker task requires a structured caption $\hat{y} = \{\hat{y}_d\}_{d\in\mathcal{D}}$, where $\mathcal{D}$ contains 12 structured dimensions and one free-form summary. The caption describes how the voice sounds and how speech is delivered. For multi-speaker audio, the model returns

$$\hat{Y} = \{(\hat{s}_k, \hat{\mathcal{I}}_k, \hat{c}_k)\}_{k=1}^{K}, \qquad \hat{\mathcal{I}}_k = \{(\hat{t}^0_{kj}, \hat{t}^1_{kj})\}_{j=1}^{J_k},$$

where K is the predicted number of speakers, $\hat{s}_k$ is a within-recording speaker ID, $\hat{\mathcal{I}}_k$ contains all predicted activity intervals for that speaker, and $\hat{c}_k$ is a caption using the same 13-field schema. Each speaker is represented by one item, even when active in multiple intervals.

The benchmark contains 1,000 single-speaker clips and 1,000 controlled multi-speaker mixtures. The single-speaker set comprises 480 Chinese clips and 520 non-Chinese clips, including 454 English clips and 66 in other languages. Single-speaker durations range from 3 to 30 s, with 87% between 3 and 15 s (Appendix A).

3.2 Caption Schema

Table 1 organizes 12 vocal attributes into four groups: identity presentation, prosody and acoustics, articulation, and affect and style. These groups consolidate dimensions used in prosody, speaking-style description, and controllable speech (Cutler et al., 1997; Jin et al., 2024; Diwan et al., 2025; Huang et al., 2025b). A free-form `summary` integrates the most salient cues, yielding 13 fields in total. Identity-related fields and `personality` record perceived vocal presentation and impressions supported by the clip.

Table 1: Structured caption schema of VocalCap: twelve vocal attributes in four groups, plus a free-form summary. The identity fields describe perceived acoustic presentation, not verified personal identity.

Dimension	Description
Speaker identity presentation	
gender	Gender presentation inferred from voice
age	Life stage suggested by voice
accent	Accent, dialect, or non-native pattern
Prosody / acoustics	
pitch	Perceived pitch range and movement
volume	Loudness, whispering, or shouting
speed	Speaking rate and rhythmic style
texture	Timbre, e.g., bright, gravelly, nasal
Articulation	
clarity	Pronunciation clarity, mumbling, slurring
fluency	Hesitation, stuttering, filler-heavy delivery
Affect / style	
emotion	Expressed affect, e.g., calm, angry
tone	Attitudinal quality, e.g., sincere, sarcastic
personality	Perceived impression, e.g., confident, anxious
summary	Free-form description of salient vocal properties

Multi-speaker references use the same field definitions and inherit descriptions from the source clips. Their fidelity after acoustic mixing is assessed in Appendix C, including fields whose perception can change with gain, noise, or overlap.

3.3 Source Selection and Annotation

We select acted film and television speech for its expressive vocal delivery. Candidates pass six cleaning stages: voice-activity filtering, single-speaker verification, ASR validity checking, audio-quality filtering, overlap filtering, and duration control (Figure 2). Diarization and speaker verification check for a single consistent source, while Whisper provides the ASR validity check (Radford et al., 2023). Accepted clips have a speech ratio of at least 0.6, DNSMOS of at least 3.0, and SNR of at least 15 dB. Cleaning settings are provided in Appendix A.2.

Fifteen trained annotators describe the 13 fields using dimension-specific guidelines and flag uncertain judgments for review. Two reviewers re-annotate a random sample of 100 clips, covering 10% of the single-speaker set. The reported pooled agreement is $\kappa = 0.79$ (95% CI $[0.76, 0.82]$; 1,200 clip–attribute decisions). Appendix A.3 reports annotation and agreement details. Disagreements are reviewed before references enter evaluation, and unresolved reference fields are excluded from scoring. Table S3 additionally reports per-field agreement. The release provides annotations, source pointers, timestamps, and construction metadata; the underlying film and television audio is not redistributed (Appendix H).

3.4 Tag Space and Balanced Sampling

To control attribute coverage, we map human-written reference phrases to canonical sampling tags. Compact fields use predefined tags. For open fields, LLM-assisted taxonomy induction operates on the reference phrases under a fixed category budget for each dimension, followed by human-audited mapping. The full taxonomy defines 167 leaf tags; 145 occur in the released set (Appendix B). Each record retains both its original phrase and a sampling tag. Evaluation uses the phrase, preserving details such as pitch movement or multiple timbral cues that a single tag may omit.

Selection proceeds in two stages. Inverse-frequency sampling first upweights clips containing rare tags. Greedy histogram matching then reduces deviations from target tag counts across all 12 marginal distributions, followed by local swaps that further improve the match (Appendix B). These

operations select one shared set of 1,000 clips for all dimensions. The `summary` field is evaluated but does not define sampling buckets.

4 MULTI-SPEAKER TRACK

The multi-speaker track extends the 13-field caption schema to recordings in which models must identify speakers, localize their activity, and associate each caption with the correct speaker. We construct 1,000 mixtures from 735 eligible clips in the single-speaker set, retaining source correspondence and activity intervals for evaluation.

Controllable synthesis. Each mixture contains two to six speakers arranged in one of four interaction patterns: turn-taking, partial overlap, foreground/background, or speaker dominance. The generator varies segment duration, inter-segment gaps, speaker roles, and speaking-time balance. Positive gaps create silence, near-zero gaps produce adjacent turns, and negative gaps introduce overlap. Speaker selection favors acoustically distinguishable voices while including harder combinations with similar timbre or pitch. Target difficulty guides overlap, speaker similarity, and noise settings. Our reported difficulty analysis uses observed speaker counts and overlap bins. The released mixtures average 9.9 s (range 3.0–22.9 s) and 3.06 speakers. The mean overlap ratio is 0.11, with a cap near 0.45; at most two speakers are simultaneously active in 95% of mixtures.

Acoustic mixing and metadata. The mixer adjusts segment boundaries toward low-energy regions, applies fades, and sets per-speaker gain from -9.6 to $+4.0$ dB. Noisy mixtures use target SNRs of 9, 16, or 25 dB; clean mixtures contain no added noise. Audio is rendered as 16 kHz mono FLAC without artificial reverberation. Table S5 distinguishes generation settings from realized statistics. Gain or SNR can vary with a fixed timeline; varying overlap requires changing offsets and updating reference intervals. The reported slices compare existing mixtures rather than matched re-renderings. Metadata include RTTM-style speaker segments, word timestamps, language, speaker IDs, source slices, and the 13-field captions inherited from the source clips. These captions remain source-level references: gain, noise, and overlap can alter perceived loudness, clarity, texture, or emotional salience. Source eligibility criteria, parameter ranges, composition counts, and a fidelity audit of 100 rendered speaker segments are provided in Appendix C.

5 EVALUATION

5.1 EVALUATION PROTOCOL

Both tracks use the same 13-field caption schema. A multi-speaker output additionally assigns each caption a within-recording speaker ID and activity intervals. We remove formatting wrappers, recover unambiguous fields, canonicalize IDs, and clip intervals to the recording duration. For mixtures, the parsed-output rate measures recovery of a nonempty speaker inventory. Empty or unrecoverable inventories receive zero components and remain in evaluation.

Qwen3-235B scores each field against its human reference using deterministic decoding. Each call contains the field definition, reference, and prediction, without audio or system identity. Speaker correspondence is determined solely from temporal information. Inference and judging prompts and scoring rules are provided in Appendix D; calibration is described in Appendix E.

5.2 SINGLE-SPEAKER CAPTIONING

The judge scores each field from 1 to 5. We average scored samples within each field, then take the unweighted mean across the 13 fields to obtain S and normalize it as $\tilde{S} = 100(S-1)/4$. Outputs that cannot be parsed into the schema are excluded from that system's means; excluded counts are reported in Table S9.

5.3 MULTI-SPEAKER EVALUATION

We report Coverage, Timeline, and Binding on $[0, 1]$. All three can be affected by missing speakers and should be interpreted jointly.

Coverage. We merge each speaker's overlapping intervals into a union of active time and compute temporal IoU between the reference and predicted unions. One-to-one matching maximizes total IoU over pairs with IoU ≥ 0.5; a greedy fallback applies if either inventory exceeds 20 speakers (Appendix D.3). With G reference speakers, P predictions, and m accepted matches, precision is m/P, recall is m/G, and Coverage is their harmonic mean:

$$\text{Coverage} = \frac{2m}{G+P}. \tag{1}$$

Coverage is zero when no pair is accepted.

Timeline. We convert activity intervals to a 10-ms binary grid and exclude frames within $0.25\,\text{s}$ of any reference boundary. A separate one-to-one assignment maximizes overlapping active frames without an IoU threshold. Each mapped reference receives its frame-level Jaccard error; unmapped references receive error 1. Their reference-speaker mean is the Jaccard error rate (JER), giving

$$\text{Timeline} = 1 - \text{JER}. \tag{2}$$

Binding. We score the 13 caption fields of each Coverage-matched pair on a 1–10 rubric. Let $s_g = 13^{-1}\sum_{d=1}^{13} s_{g,d}$ average all 13 valid field scores for a matched reference speaker g, with $b_g = (s_g - 1)/9$. Unmatched references have $b_g = 0$. Failed or missing judge responses must be completed before aggregation; they are not model errors or rubric-floor scores. Binding retains every reference in its denominator:

$$\text{Binding} = \frac{1}{G}\sum_{g=1}^{G} b_g. \tag{3}$$

Empty or abstaining field predictions receive the rubric floor. Score reuse and boundary cases are specified in Appendix D.3.

Overall score. We combine the components within each mixture before averaging:

$$\text{Multi} = \frac{100}{N}\sum_{i=1}^{N}\left(\text{Coverage}_i\,\text{Timeline}_i\,\text{Binding}_i\right)^{1/3}. \tag{4}$$

All $N = 1{,}000$ mixtures enter the aggregate. We report percentile 95% CIs from 2,000 mixture bootstrap resamples. Full-span controls predict three speakers or the reference count, each active throughout the recording; their empty captions give zero Binding and Multi.

Diagnostic evaluations. On 200 stratified mixtures, we randomly reassign complete captions among predicted speakers so that none retains its original assignment, keeping the inventory, intervals, and temporal matches fixed. The margin is original minus shuffled Binding. Mixtures with fewer than two predictions are excluded; system-specific effective counts and standard errors are reported in Appendix E.2. Within each difficulty group, recall and precision use all mixtures; T_{cond} measures frame-level Jaccard for Coverage pairs using the same grid and collars. We average it within each mixture, then over mixtures with accepted matches, reporting their counts. GT-assisted runs receive reference speaker IDs and intervals as prompt inputs; scoring uses their actual outputs.

Unified benchmark score. For systems evaluated on both tracks, we report $\text{VC} = 0.4\tilde{S} + 0.6\tilde{M}$, where $\tilde{M} = \text{Multi}$ is already on $[0, 100]$. Both track scores accompany VC; weight sensitivity and alternative aggregations are reported in Appendix E.4.

6 Experiments

Models and evaluation. We evaluate 14 systems on the single-speaker track and 12 on the multi-speaker track; the latter excludes Gemini-3-Pro and Gemini-3.6-Flash. The system inventory and documented inference settings are in Appendix D. Qwen3-235B compares predicted fields with human references using 1–5 and 1–10 rubrics for the respective tracks, without access to audio or system identity. On seven open single-speaker fields, judge–human Spearman correlation is $\rho = 0.79$ and mean absolute deviation is 0.40; matched and mismatched probes receive scores in the expected direction in 88% of cases (Appendix E).

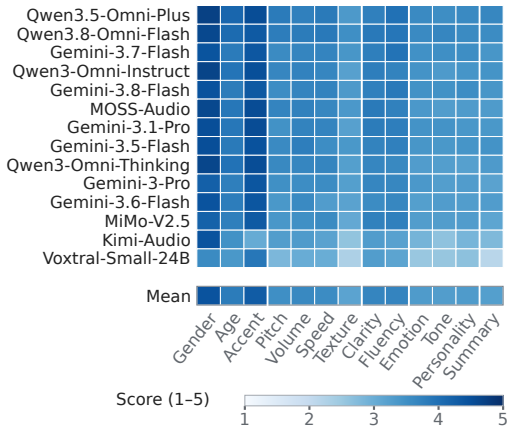

Figure 3: Single-speaker field scores for 14 systems on 1,000 clips (1–5 scale). Rows follow overall score; Mean is the unweighted system average. Unparseable outputs are excluded; counts are in Table S9.

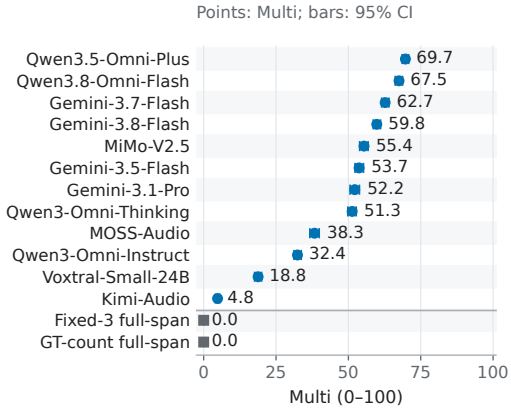

Figure 4: Multi scores on 1,000 mixtures with percentile 95% CIs from 2,000 mixture bootstrap resamples (Section 5.3). Parsing failures receive zero scores and remain in evaluation. Full-span controls use empty captions.

6.1 Single-Speaker Results

Figure 3 shows the field profiles of all 14 systems. Qwen3.5-Omni-Plus leads overall with a normalized score of 72.2, followed by Qwen3.8-Omni-Flash at 71.9. Field-level leaders vary: Qwen3.8-Omni-Flash leads on volume, speed, texture, clarity, and tone, while Gemini-3.1-Pro leads on accent.

Gender and accent generally receive higher reference agreement than texture, emotion, tone, and personality. For Qwen3.5-Omni-Plus, gender scores 4.748, compared with 3.362 for texture and 3.550 for tone. These differences reveal uneven attribute profiles, interpreted alongside human annotation agreement in Section 7. Single-speaker slice results are reported in Appendix F.

6.2 Multi-Speaker Results

Qwen3.5-Omni-Plus also leads the multi-speaker track with 69.7 (95% CI $[68.6, 70.8]$), followed by Qwen3.8-Omni-Flash at 67.5 (Figure 4). Across the 12 shared systems, Spearman rank correlation is $\rho = 0.727$, but 15 of the 66 system pairs reverse order between tracks. For example, Qwen3-Omni-Instruct moves from fourth to tenth. The leader's Coverage, Timeline, and Binding are 0.855, 0.814, and 0.497. Because matching failures affect these components jointly, we next examine caption assignment and structural assistance. Full component profiles and unified scores are in Appendix F.

6.3 Speaker-Grounded Diagnostics

Caption assignment affects Binding. On 200 stratified mixtures, we fix each system's predicted inventory and intervals and reassign complete captions so that none remains with its original speaker. Coverage and Timeline therefore remain unchanged. For Qwen3.8-Omni-Flash, Binding falls from 0.507 to 0.302 on 198 eligible mixtures, giving a margin of 0.205 (SE 0.009; Figure 5). This shows that the original assignments better preserve speaker-specific reference agreement. Effective subsets and baseline Binding differ across systems; the margins are within-system diagnostics. Results and the derangement protocol are in Appendix E.2.

GT structure improves the complete output. We compare Qwen3.8-Omni-Flash under free generation and with reference speaker IDs and activity intervals supplied in the prompt. On the same 200 mixtures, Multi increases from 69.6 to 80.4, a gain of 10.8 points (Table 2). The assisted condition recovers a nonempty inventory in 195 cases and preserves the complete GT structure in 192; all 200 enter evaluation. The Binding increase also reflects improved matching coverage.

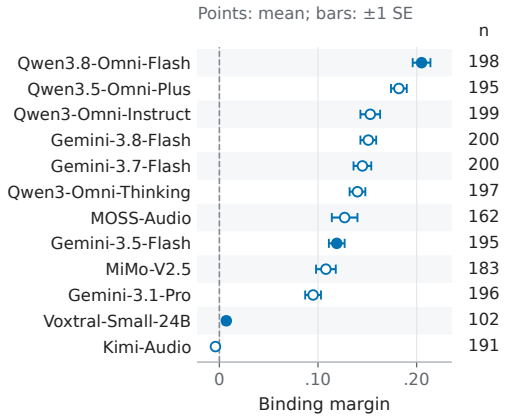

Figure 5: Original minus shuffled Binding; bars show ±1 SE across eligible mixtures. Right: effective n. Subsets differ across systems; source reuse is not reflected in the SEs.

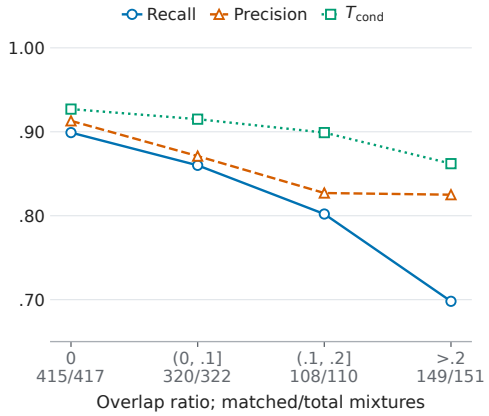

Figure 6: Qwen3.5-Omni-Plus across overlap bins. Recall and precision average over all mixtures; T_{cond} uses mixtures with Coverage matches. Counts show conditional/total mixtures. Points are group means without CIs; bin composition differs.

Table 2: Qwen3.8-Omni-Flash on the same 200 mixtures. Both conditions use the same judge and include failures. The assisted prompt provides the GT speaker inventory and intervals; matched-pair field scores are complete in both conditions.

Condition	C	T	B	Multi
Free generation	0.851	0.803	0.503	69.6
GT inventory and timeline	0.969	0.967	0.560	80.4

6.4 Multi-Speaker Difficulty Analysis

We examine matching coverage alongside T_{cond}, frame-level Jaccard on Coverage-matched speakers, averaged first within a mixture and then over mixtures with accepted matches (Appendix D.4). For Qwen3.5-Omni-Plus, recall decreases from 0.899 without overlap to 0.698 above 0.2 overlap, while T_{cond} decreases from 0.927 to 0.862 (Figure 6). Conditional statistics cover 415/417 and 149/151 mixtures, respectively. Lower recall can reflect omitted speakers or intervals that fail the matching threshold.

Speaker-count slices show why composition matters: recall falls from 0.897 for two speakers to 0.776 for five, then rises to 0.877 for six. The six-speaker group contains only 19 mixtures, all without overlap. Conditional caption scores vary only from 6.275 to 6.358 across overlap bins (shown in Appendix F), but exclude unmatched speakers, highlighting why descriptive agreement must be interpreted alongside speaker recall.

6.5 Benchmark Validation

Sensitivity to score aggregation. Using the reported track means, we recompute $\text{VC}_w = w\tilde{S} + (1-w)\tilde{M}$ for $w \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$. The first four systems retain their order at every setting, and Kendall's τ with the default $w = 0.4$ ranking ranges from 0.939 to 1.000. Geometric and harmonic means reproduce the full default ordering of all 12 systems (Appendix E.4). Thus, the leading positions persist across these aggregation choices, while some middle ranks depend on the track weights. This analysis changes the combination of track scores, leaving the per-mixture definition of Multi fixed.

Reference fidelity after mixing. The manual audit of 100 rendered speaker segments reports group-level source-annotation retention rates of 99%, 95%, and 90% for the low-, moderate-, and high-sensitivity attribute groups, respectively (Appendix C). The groups include gender, age, and accent; volume, personality, and coarse texture; and clarity, fine texture, and emotion, respectively. These rates summarize how often the sampled source descriptions remained applicable after rendering.

7 Discussion and Limitations

The diagnostics indicate that descriptive agreement should be interpreted jointly with speaker recovery and temporal grounding.

Human agreement on the source captions is higher for gender and accent ($\kappa = 0.91$ and 0.83) than for texture, emotion, tone, and personality (0.41–0.52; Appendix A.3). Lower model scores on the latter fields should therefore be interpreted alongside annotation variability.

Scope of the mixtures. The controlled mixtures provide traceable source identities and intervals, but average only 9.9 s; 95% contain at most two simultaneously active speakers (Appendix C). They offer limited evidence about long conversations, sustained dense overlap, and context-dependent interaction. Predominantly Chinese and English acted speech further limits generalization to spontaneous dialogue and other languages. Reused source clips also induce dependence between mixtures that the reported mixture-level uncertainty estimates do not capture.

Reference and judge limits. Source captions can change in applicability after acoustic mixing: the 100-segment audit reports 90% retention for the most rendering-sensitive group (Appendix C). The text-only judge measures reference agreement without establishing which acoustic or lexical cues a model used. Multi-speaker references inherit the reviewed source captions, whose annotation agreement is reported above (Appendix A.3). A secondary DeepSeek-V4-Pro judge largely preserves single-speaker ordering (Spearman $\rho = 0.952$). The multi-speaker secondary-judge evaluation covers 11 systems using the same predictions as the primary evaluation (Appendix E). Personality is scored as a perceived vocal impression supported by the available evidence.

Causal interpretation. Caption shuffling isolates reassignment within predicted structures, but its margins depend on coverage, caption similarity, and eligible subsets. GT assistance changes both speaker-inventory and timeline information and has been evaluated on one model; its gain does not isolate timestamp accuracy or define an oracle upper bound. Difficulty slices vary in composition, while conditional metrics exclude unmatched speakers. Matched acoustic interventions and enforced reference intervals would further clarify error sources.

Data and ethics. Gender, age, accent, and personality describe perceived vocal presentation and impressions, not verified identity or psychological traits. VocalCap should not support inferences about protected identity or mental state. Raw film and television audio is not redistributed; annotations, source pointers, and construction code support reconstruction from legally held copies. Source rights, annotation subjectivity, and distribution are detailed in Appendix H.

8 Conclusion

VocalCap connects fine-grained vocal description with speaker recovery and temporal localization across single-speaker and controlled multi-speaker tracks. A shared 13-field schema supports comparison, while Coverage, Timeline, and Binding assess the completeness and correspondence of predicted outputs. Cross-track rank changes show that performance on isolated voices does not fully predict performance on complete mixtures. Caption shuffling lowers Binding for most systems, while reference speaker IDs and intervals improve Qwen3.8-Omni-Flash's Multi score by 10.8 points on 200 mixtures. For Qwen3.5-Omni-Plus, greater overlap accompanies lower matching recall and conditional temporal accuracy. These findings motivate evaluating descriptions jointly with speaker assignments and activity timelines, and extending this controlled evaluation to natural dialogue and sustained overlap.

AI USE STATEMENT

LLMs assist taxonomy induction and mapping of human-written reference phrases to sampling tags (Appendix B); they do not originate per-clip references. Qwen3-235B serves as a text-only evaluation judge, with its documented protocol and calibration scope in Appendices D and E. Audio-language models are the evaluated systems. An AI assistant was used to revise prose, organize appendices, check reported numerical consistency, and update LaTeX and schematic figures.

ETHICS STATEMENT

VocalCap describes acoustic presentation and perceived expressive qualities, not verified personal identity or psychological traits. Acted speech and subjective references limit generalization; text-only judging measures agreement with the human reference descriptions (Section 7). Raw film and television audio is not redistributed. Source rights, release scope, and responsible use are discussed in Appendix H.

REPRODUCIBILITY STATEMENT

Appendices A–C describe data, annotation, sampling, and mixture construction. Appendix D records the API model strings, complete inference prompts, parsing, matching, and scoring conventions. Appendices E and F give judge validation, secondary-judge checks, diagnostics, and numerical results. Appendix H summarizes the released annotations, source pointers, configuration metadata, and evaluation assets needed to reproduce the reported snapshot.

REFERENCES

Atsushi Ando, Takafumi Moriya, Shota Horiguchi, and Ryo Masumura. Factor-conditioned speaking-style captioning. In *Interspeech 2024*, pp. 782–786, 2024. doi: 10.21437/Interspeech.2024-633. URL https://doi.org/10.21437/interspeech.2024-633.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. pyannote.audio: Neural building blocks for speaker diarization. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7124–7128, 2020. doi: 10.1109/ICASSP40776.2020.9052974. URL https://doi.org/10.1109/ICASSP40776.2020.9052974.

CHEN CHEN, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and Ensiong Chng. Audio large language models can be descriptive speech quality evaluators. In *International Conference on Learning Representations*, volume 2025, pp. 24920–24934, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3e523b2c0ab0783b753ad681ea39cf32-Paper-Conference.pdf.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024. URL https://arxiv.org/abs/2407.10759.

Anne Cutler, Delphine Dahan, and Wilma van Donselaar. Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40(2):141–201, 1997. doi: 10.1177/002383099704000203. URL https://doi.org/10.1177/002383099704000203.

Anuj Diwan, Zhisheng Zheng, David Harwath, and Eunsol Choi. Scaling rich style-prompted text-to-speech datasets. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3639–3659, 2025. doi: 10.18653/v1/2025.emnlp-main.180. URL https://aclanthology.org/2025.emnlp-main.180/.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740, 2020. doi: 10.1109/ICASSP40776.2020.9052990. URL https://doi.org/10.1109/ICASSP40776.2020.9052990.

Jonathan G. Fiscus, Jerome Ajot, Martial Michel, and John S. Garofolo. The rich transcription 2006 spring meeting recognition evaluation. In *Machine Learning for Multimodal Interaction*, pp. 309–322. Springer Berlin Heidelberg, 2006. doi: 10.1007/11965152_28. URL https://doi.org/10.1007/11965152_28.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Evuru, Ramaneswaran S, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Compa: Addressing the gap in compositional reasoning in audio-language models. In *International Conference on Learning Representations*, volume 2024, pp. 15486–15511, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/43c18853329c7504996b255252b6cb1f-Paper-Conference.pdf.

Chien-yu Huang, Min-Han Shih, Ke-Han Lu, Chi-Yuan Hsiao, and Hung-yi Lee. Speechcaps: Advancing instruction-based universal speech models with multi-talker speaking style captioning. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025a. doi: 10.1109/ICASSP49660.2025.10889205. URL https://doi.org/10.1109/ICASSP49660.2025.10889205.

Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, Shimin Li, Zhaoye Fei, Qinyuan Cheng, and Xipeng Qiu. Instructttseval: Benchmarking complex natural-language instruction following in text-to-speech systems. *arXiv preprint arXiv:2506.16381*, 2025b. URL https://arxiv.org/abs/2506.16381.

Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu. Speechcraft: A fine-grained expressive speech dataset with natural language description. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1255–1264, 2024. doi: 10.1145/3664647.3681674. URL https://doi.org/10.1145/3664647.3681674.

Masaya Kawamura, Ryuichi Yamamoto, Yuma Shirahata, Takuya Hasumi, and Kentaro Tachibana. Libritts-p: A corpus with speaking style and speaker identity prompts for text-to-speech and style captioning. In *Interspeech 2024*, pp. 1850–1854, 2024. doi: 10.21437/Interspeech.2024-692. URL https://doi.org/10.21437/interspeech.2024-692.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011/.

Ruohan Liu, Shukang Yin, Tao Wang, Dong Yan Zhang, Weiji Zhuang, Shuhuai Ren, Ran He, Caifeng Shan, and Chaoyou Fu. Speechparaling-bench: A comprehensive benchmark for paralinguistic-aware speech generation. *arXiv preprint arXiv:2604.20842*, 2026. URL https://arxiv.org/abs/2604.20842.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman. The second dihard diarization challenge: Dataset, task, and baselines. In *Interspeech 2019*, pp. 978–982, 2019. doi: 10.21437/Interspeech.2019-1268. URL https://doi.org/10.21437/interspeech.2019-1268.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d36f208919582785db965fe648b9fe59-Paper-Conference.pdf.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 16607–16629, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/476ab8f369e489c04187ba84f68cfa68-Paper-Conference.pdf.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4297–4316, 2025a. doi: 10.18653/v1/2025.naacl-long.218. URL https://aclanthology.org/2025.naacl-long.218/.

Helin Wang, Jiarui Hai, Dading Chong, Karan Thakkar, Tiantian Feng, Dongchao Yang, Junhyeok Lee, Thomas Thebaud, Laureano Moro-Velazquez, Jesus Villalba, Zengyi Qin, Shrikanth Narayanan, Mounya Elhiali, and Najim Dehak. Capspeech: Enabling downstream applications in style-captioned text-to-speech. *IEEE Transactions on Audio, Speech and Language Processing*, 34:3519–3531, 2026. doi: 10.1109/TASLPRO.2026.3703247. URL https://doi.org/10.1109/TASLPRO.2026.3703247.

Qiongqiong Wang, Hardik Bhupendra Sailor, Tianchi Liu, Wenyu Zhang, Muhammad Huzaifah, Nattadaporn Lertcheva, Shuo Sun, Nancy F. Chen, Jinyang Wu, and AiTi Aw. Benchmarking contextual and paralinguistic reasoning in speech-LLMs: A case study with in-the-wild data. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14133–14148, 2025b. doi: 10.18653/v1/2025.findings-emnlp.760. URL https://aclanthology.org/2025.findings-emnlp.760/.

Kazuki Yamauchi, Yusuke Ijima, and Yuki Saito. Stylecap: Automatic speaking-style captioning from speech based on speech and language self-supervised learning models. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11261–11265, 2024. doi: 10.1109/ICASSP48485.2024.10445977. URL https://doi.org/10.1109/ICASSP48485.2024.10445977.

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, Zhengyuan Lin, Ziqi Chen, Zhaoye Fei, Chenghao Liu, Donghua Yu, Jun Zhan, Kang Yu, Kexin Huang, Liwei Fan, Mingshu Chen, Qinyuan Cheng, Ruixiao Li, Shimin Li, Songlin Wang, Xingjian Zhao, Yang Gao, Yitian Gong, Yiyang Zhang, Zhe Xu, and Xipeng Qiu. Moss-audio technical report. *arXiv preprint arXiv:2606.01802*, 2026a. URL https://arxiv.org/abs/2606.01802.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, 2024. doi: 10.18653/v1/2024.acl-long.109. URL https://aclanthology.org/2024.acl-long.109/.

Shu-wen Yang, Ming Tu, Ting-Wei Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. Paras2s: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. In *International Conference on Learning Representations*, volume 2026, pp. 70704–70729, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/734b2e77222680728b9ce78c573eae1e-Paper-Conference.pdf.

APPENDIX

Appendices A–C document the benchmark data and construction; Appendix D describes the evaluation protocol. Judge validation and diagnostic controls appear in Appendix E, full track results and difficulty slices in Appendix F, related task interfaces in Appendix G, and release scope in Appendix H.

A BENCHMARK COMPOSITION AND HUMAN ANNOTATION

A.1 SOURCES AND COMPOSITION

The single-speaker track contains one shared set of 1,000 clips, with all 13 fields annotated on each clip. Chinese film and television contribute 480 clips; foreign television series contribute 520, including 454 English and 66 other-language clips. These are expressive acted recordings, not a representative sample of everyday dialogue. Table S1 provides the consolidated statistics. Source IDs and timestamps support auditing without publishing named speaker labels.

Table S1: Benchmark composition. The 12 structured attributes share the same 1,000 single-speaker clips; the summary adds the thirteenth field.

Statistic	Value
Single-speaker clips	1,000
Chinese / English / other-language clips	480 / 454 / 66
Duration range; fraction within 3–15 s	3–30 s; 87%
Median / mean duration	6.8 / 8.82 s
Structured attributes / total caption fields	12 / 13
Defined / observed canonical tags	167 / 145
Eligible single-speaker sources used for mixtures	735
Multi-speaker mixtures	1,000
Speakers per mixture, range / mean	2–6 / 3.06

A.2 CLEANING AND ELIGIBILITY

Candidate cleaning includes voice-activity filtering, single-speaker verification, ASR sanity checks, audio-quality filtering, overlap filtering, and duration control. Table S2 lists the documented criteria. Retained clips span 3–30 s, with 87% between 3 and 15 s. The construction summary does not record the historical duration-filter settings, exception rules, or exact tool versions.

Table S2: Documented cleaning criteria and retained-set duration statistics. The duration row describes the retained clips, rather than a filtering threshold.

Stage	Tool	Reported criterion
Voice activity	VAD	Speech ratio ≥ 0.6
Single-speaker	Diarization / verification	One speaker, no change point
ASR sanity	Whisper (Radford et al., 2023)	Nonempty, valid text
Audio quality	DNSMOS / SNR	DNSMOS ≥ 3.0; SNR ≥ 15 dB
Overlap	Overlap detector	Overlap ratio ≤ 0.1
Duration	Retained-set statistic	3–30 s; 87% within 3–15 s

A.3 ANNOTATION AND AGREEMENT

Fifteen trained annotators produce field-specific descriptions and flag uncertain judgments. Two reviewers re-annotate a random 100-clip sample. Table S3 reports the agreement statistics. The pooled statistic is based on 1,200 clip–attribute decisions and is not the arithmetic mean of the per-field κ values. These agreement estimates characterize the source-caption annotations, which are inherited by the multi-speaker references. They do not measure human agreement on mixture-level

speaker assignment or interval quality. The phrase-coding scheme, treatment of the two reviewers, and timing of agreement calculation relative to adjudication are not documented with these statistics.

Table S3: Per-dimension agreement between the initial annotation and reviewer re-annotation on the 10% quality-control sample ($n = 100$ clips). Weighted κ is additionally reported for ordered dimensions. The pooled row contains 1,200 clip–dimension decisions.

Dimension	Cohen's κ	95% CI	Weighted κ	n
gender	0.91	$[0.84, 0.97]$	—	100
accent	0.83	$[0.75, 0.90]$	—	100
age	0.71	$[0.61, 0.80]$	0.82	100
clarity	0.72	$[0.62, 0.81]$	—	100
speed	0.69	$[0.59, 0.78]$	0.78	100
volume	0.67	$[0.57, 0.77]$	0.76	100
fluency	0.66	$[0.55, 0.76]$	—	100
pitch	0.63	$[0.52, 0.73]$	0.74	100
texture	0.52	$[0.40, 0.63]$	—	100
tone	0.48	$[0.36, 0.60]$	—	100
emotion	0.46	$[0.34, 0.58]$	—	100
personality	0.41	$[0.29, 0.53]$	—	100
Pooled (12 dimensions)	0.79	$[0.76, 0.82]$	—	1,200

Disagreements are reviewed before references enter evaluation. Unresolved reference fields are excluded from scoring. By contrast, an empty, unknown, or abstaining model prediction receives the rubric floor; it is not a failed judge call. The number of unresolved reference fields per dimension is not recorded in the available summary.

B Taxonomy and Balanced Sampling

B.1 Reference phrases and canonical tags

Human-written phrases remain the semantic evaluation references. Each structured field also stores one canonical tag for sampling and analysis. A phrase may describe multiple cues, such as pitch movement or several timbral qualities, that its single tag does not preserve. Approximately 0.9% of mapped fields use a fallback bucket without replacing their reference phrases.

Compact fields use predefined categories. For open fields, LLM-assisted induction processes phrases in descending frequency order under a per-dimension category budget, followed by human-audited mapping. Once the budget is reached, subsequent phrases map to existing categories. Categories with fewer than 100 candidate-pool occurrences are flagged as low-resource. The full vocabulary has 167 tags; Table S4 lists the 145 observed tags across dimensions in the selected set. Tag counts measure coverage, not closeness to the target histograms.

Table S4: Distinct tags observed in the same 1,000 clips. Language composition and sample count are shared across all rows (Table S1). The summary field has no sampling bucket.

Field	Tags	Field	Tags	Field	Tags
gender	3	volume	5	fluency	9
age	7	speed	8	emotion	24
accent	9	texture	22	tone	23
pitch	6	clarity	10	personality	19

B.2 SELECTION OBJECTIVE AND PROCEDURE

Let $\mathcal{P}$ be the cleaned pool, $t_d(a)$ a clip's tag in dimension d, and $\hat{p}_d(\mathcal{P})$ the corresponding pool marginal. Inverse-frequency weighting favors rare combinations using

$$s(a) = -\sum_{d=1}^{12} \log \hat{p}_d(\mathcal{P})[t_d(a)], \qquad w(a) = \exp s(a). \tag{5}$$

For sequential weighted draws without replacement, the normalized weight describes the *next draw* from the remaining pool $\mathcal{P}_t$:

$$\Pr(a_t = a \mid \mathcal{P}_t) = \frac{w(a)}{\sum_{a' \in \mathcal{P}_t} w(a')}. \tag{6}$$

It is not a clip's final inclusion probability after multiple draws.

The subsequent histogram-matching stage seeks one shared subset with target size $K = 1{,}000$. Writing $n_d(S)$ for tag counts in S and τ_d for target counts, its objective is

$$\min_{S \subseteq \mathcal{P},\, |S|=K} \mathcal{L}(S), \qquad \mathcal{L}(S) = \sum_{d=1}^{12} \|n_d(S) - \tau_d\|_2^2. \tag{7}$$

The documented greedy procedure shuffles candidates, accepts additions that reduce $\mathcal{L}$ (with a small random-acceptance probability ϵ), and uses improving swaps for local refinement. Its objective is approximate matching of the twelve marginal histograms. The selected set contains 1,000 clips. The construction summary does not specify the target histograms, random seed, ϵ, the procedure for reaching 1,000 clips if candidates are exhausted, or the final histogram deviations.

C CONTROLLED MULTI-SPEAKER CONSTRUCTION

C.1 SOURCE ELIGIBILITY, TIMELINES, AND RENDERING

Mixtures use 735 eligible sources from the single-speaker set. The remaining candidates are filtered for ASR speech windows shorter than 2.5 s, fewer than eight aligned CTC tokens, ambiguous gender labels, or non-speech/dirty tags. ASR/CTC-guided foreground slices contain at least six tokens and short responses at least four; slices reused from a source clip are separated by at least 0.25 s. These criteria apply to source eligibility and slicing, separately from the single-speaker cleaning thresholds.

The generator selects speaker count, interaction pattern, target difficulty, attribute contrast, and noise condition, then places source slices on a shared timeline. Table S5 separates settings from observed mixture statistics. Gain and noise can change while source intervals remain fixed. Changing overlap requires moving intervals, while retaining source content and turn order where possible, and regenerating reference timestamps. These are generator capabilities; the reported difficulty slices are observational group comparisons, not matched acoustic interventions.

Table S5: Generator settings. Realized duration, overlap, and concurrency statistics are reported separately.

Control	Setting or reported range
Interaction pattern	Turn-taking, partial overlap, foreground/background, dominance
Speaker count	2–6
Segment duration	1.5–9.0 s
Inter-segment gap	-0.9–0.8 s; negative gaps introduce overlap
Speaking-time balance	Background duration 0.4–$0.9\times$ foreground
Per-speaker gain	-9.6–$+4.0$ dB
Noise target SNR	9, 16, or 25 dB; clean condition has no added noise
Speaker similarity	Easy/hard selection categories
Output	16-kHz mono FLAC; no artificial reverberation

C.2 REALIZED COMPOSITION AND SOURCE REUSE

The 1,000 mixtures average 9.9 s (3.0–22.9 s) and 3.06 speakers. We define the recorded overlap ratio r as the duration with at least two active speakers divided by the union of all active speech time; the reported bins are $r = 0$, $0 < r \leq 0.1$, $0.1 < r \leq 0.2$, and $r > 0.2$. Mean r is 0.11, with a cap near 0.45; 95% contain at most two simultaneously active speakers. Table S6 records the realized speaker-count, interaction, acoustic, language, and attribute-contrast strata. Speaker-similarity labels are generator controls used during source selection, whereas the results below use the observed speaker-count and overlap metadata. Aggregate easy/medium/hard labels are not used as an additional score axis.

Sources recur across mixtures. Consequently, mixture-level bootstrap intervals and diagnostic standard errors do not account for dependence induced by source reuse. They describe variation over the evaluated mixtures rather than over independent source speakers.

Table S6: Composition of the released multi-speaker set (1,000 mixtures). Speaker counts average 3.06, matching the ground-truth count reported in the main paper.

Category	N
Speaker count	
2 / 3 / 4 / 5 / 6	394 / 306 / 164 / 117 / 19
Interaction pattern	
turn-taking	256
partial overlap	363
foreground/background	182
speaker dominance	199
Acoustic condition	
clean	550
light environment noise	276
low-volume music / natural	80
background human/env noise	76
real music	18
Language	
zh / en / en+zh	461 / 395 / 144
Attribute contrast	
gender / pitch / emotion	200 / 200 / 200
speed / style-texture / similar-hard	150 / 150 / 100

C.3 SOURCE-REFERENCE FIDELITY AFTER MIXING

Each rendered speaker inherits the source clip's human-written caption and has known source correspondence and activity intervals. Gain, noise, overlap, and slicing can change the applicability of the inherited description. Table S7 summarizes retention for the listed sensitivity groups. The reported groups distinguish coarse and fine texture and do not include tone or summary.

Table S7: Group-level source-annotation retention in an audit of 100 rendered speaker segments.

Sensitivity	Reported fields	Retention
Low	Gender, age, accent, pitch, speed, fluency	99%
Moderate	Volume, personality, coarse texture	95%
High	Clarity, fine texture, emotion	90%

Reviewers examine rendered speaker segments from the evaluated mixture set and mark a source field as retained when its original description remains applicable after rendering. The audit does not establish reference fidelity under dense, sustained overlap.

D Model Inference and Evaluation Protocol

D.1 Systems and documented inference settings

Table S8 covers all 14 evaluated systems and their tracks. Single-speaker and free multi-speaker runs receive audio and a captioning instruction; reference speaker IDs and intervals are supplied only in the GT-assisted diagnostic.

Decoding uses greedy or low-temperature sampling at the settings exposed by each API (reported temperature range 0.0–0.6). The MOSS-Audio single-speaker run uses temperature 0.6, top-p 1.0, top-k 50, and 1,024 new tokens. Thinking outputs may contain reasoning before the final JSON; this is removed during parsing.

Table S8: Evaluated systems and reported identifiers. API systems list their endpoint model strings, while open-weight systems list their checkpoint identifiers. "Both" denotes the single- and multi-speaker tracks.

System	Access	Reported checkpoint / API string	Tracks
Qwen3.5-Omni-Plus	API	`qwen3.5-omni-plus`	both
Qwen3.8-Omni-Flash	API	`qwen3.8-omni-flash`	both
Gemini-3.7-Flash	API	`gemini-3.7-flash`	both
Gemini-3.8-Flash	API	`gemini-3.8-flash`	both
Gemini-3.5-Flash	API	`gemini-3.5-flash`	both
Gemini-3.6-Flash	API	`gemini-3.6-flash`	single
Gemini-3-Pro	API	`gemini-3-pro-preview`	single
Gemini-3.1-Pro	API	`gemini-3.1-pro-preview`	both
Qwen3-Omni-Instruct	Open weights	`Qwen/Qwen3-Omni-30B-A3B-Instruct`	both
Qwen3-Omni-Thinking	Open weights	`Qwen/Qwen3-Omni-30B-A3B-Thinking`	both
MiMo-V2.5	Open weights	`XiaomiMiMo/MiMo-V2.5`	both
Kimi-Audio	Open weights	`moonshotai/Kimi-Audio-7B-Instruct`	both
Voxtral-Small-24B	Open weights	`mistralai/Voxtral-Small-24B-2507`	both
MOSS-Audio	Open weights	`OpenMOSS-Team/MOSS-Audio-8B-Instruct`	both

D.2 Parsing and sample accounting

For single-speaker outputs, a lenient field-extraction parser attempts to recover the 13-field schema after removing reasoning and Markdown fences. Outputs that remain unparseable are excluded from that system's means. Table S9 gives the counts. Its retained-output count is not a guarantee of the same denominator for every field because unresolved reference fields can still affect field-specific means.

Table S9: Single-speaker parsing on 1,000 requested clips per system. Retained outputs enter field-wise evaluation; unresolved references are excluded at field level.

System	Excluded	Retained
Qwen3.5-Omni-Plus	0	1000
Qwen3.8-Omni-Flash	1	999
Gemini-3.7-Flash	0	1000
Qwen3-Omni-Instruct	0	1000
Gemini-3.8-Flash	0	1000
MOSS-Audio	8	992
Gemini-3.1-Pro	0	1000
Gemini-3.5-Flash	1	999
Qwen3-Omni-Thinking	0	1000
Gemini-3-Pro	1	999
Gemini-3.6-Flash	1	999
MiMo-V2.5	0	1000
Kimi-Audio	11	989
Voxtral-Small-24B	1	999

For multi-speaker outputs, the parser extracts a recoverable JSON array and can salvage items from surrounding text. Timestamps are in seconds and clipped to $[0, T]$; nonfinite or nonpositive-duration intervals are discarded. Items sharing a speaker ID are merged and intervals are sorted. For Coverage, overlapping or touching intervals belonging to the same speaker are merged before computing temporal intersection and union. Timeline converts activity to a binary grid. Repeated or overlapping segments therefore do not increase active duration. Unknown keys are ignored. Missing caption fields, null captions, and abstentions receive the rubric floor. Empty, missing, or unrecoverable inventories receive zero components and remain in evaluation.

The reported multi-speaker parsing rate measures recovery of a *nonempty inventory*, not strict JSON validity. A valid empty array does not count as a recovered inventory. Source/format errors are a separate bookkeeping category for missing responses, unrecoverable JSON, or upstream failures; they must not be inferred by subtracting the recovery rate from 100%. Table S18 reports recovery rates on all 1,000 mixtures.

D.3 Matching, scoring, and boundary cases

Temporal correspondence. Reference and predicted speaker IDs are canonicalized independently; valid intervals are clipped to the recording duration. For interval lists A_g and A_p, define the active-time sets $U_g = \bigcup_{a \in A_g} a$ and $U_p = \bigcup_{b \in A_p} b$. The Coverage matcher uses

$$u_{gp} = \frac{|U_g \cap U_p|}{|U_g \cup U_p|},$$

where $|\cdot|$ denotes duration and an empty union gives $u_{gp} = 0$. We sort and merge overlapping or touching intervals within each speaker before measuring these sets. Thus $u_{gp} \in [0, 1]$ and is invariant to duplicating a segment or splitting the same activity into overlapping segments. Assignment maximizes total u_{gp} over one-to-one pairs with $u_{gp} \geq 0.5$, with deterministic tie handling. If either inventory exceeds 20 speakers, the implementation instead greedily accepts eligible pairs in descending u_{gp} order without reusing a speaker; this occurs in one MiMo-V2.5 prediction containing 32 speakers. Matching never uses captions.

Frame-level timeline. For Timeline, intervals are converted to binary activity masks on a 10-ms grid, so within-speaker overlapping intervals are counted once. Start and end indices use nearest-frame rounding. We remove the union of the ± 0.25-s collars around all reference boundaries, then find a one-to-one mapping that maximizes total overlapping active frames. This mapping is separate from the thresholded Coverage assignment. A mapped reference has error $1 - |G_g \cap P_{\pi(g)}|/|G_g \cup P_{\pi(g)}|$ on retained frames; an unmapped reference has error 1. A mapped empty union has error 0. JER is the mean over reference speakers and Timeline is $1 - \mathrm{JER}$. If no reference activity remains after collar masking, Timeline is set to zero.

Caption scoring. Binding uses Qwen3-235B scores for the pairs selected by Coverage. Existing scores are reused only for the same prediction snapshot, reference, speaker pair, and field. A pair newly selected by interval-union matching requires all 13 field scores. For each matched reference g,

$$s_g = \frac{1}{13} \sum_{d=1}^{13} s_{g,d}, \qquad b_g = \frac{s_g - 1}{9}.$$

Each $s_{g,d}$ must be a successfully returned integer from 1 to 10. Failed, absent, or invalid judge responses must be resolved before aggregation; they are neither dropped from the field denominator nor replaced by score 1. A model's missing or null caption field is instead scored by the rubric. An unmatched reference has $b_g = 0$. Thus $B_i = G_i^{-1} \sum_g b_g$ retains every reference speaker. The same completion rule applies to the secondary judge and both shuffle conditions.

Aggregation and boundary cases. Each component is averaged within a mixture, then over all 1,000 mixtures, including failures. Multi is 100 times the mean of the *per-mixture* cube root $(C_i T_i B_i)^{1/3}$. The geometric mean of the three printed component means therefore does not reproduce Multi. Table S10 specifies boundary handling.

Table S10: Multi-speaker boundary cases. Component entries marked "computed" retain their deterministic definitions. Binding entries concerning one speaker describe that speaker's contribution, not necessarily the entire mixture score.

Case	C	T	B	Interpretation
Recovered inventory with scored matches	computed	computed	all-reference mean	Standard evaluation
Recovered inventory, no pair at IoU ≥ 0.5	0	computed	0	Timeline uses its separate assignment; Multi is zero
Empty, missing, or unrecoverable inventory	0	0	0	Included in all 1,000-mixture means
Unmatched reference speaker	lowers recall	zero if unmapped	zero contribution	Reference remains in the Binding denominator
Extra predicted speaker	lowers precision	computed	no extra denominator term	Accounted for through Coverage
Matched pair with 13 valid field scores	computed	computed	13-field mean	Every field has a fixed weight

Uncertainty. Percentile 95% confidence intervals use 2,000 mixture bootstrap resamples. The resampling unit is a mixture, without clustering by reused source clip.

D.4 Diagnostic execution and conditional scores

Caption shuffling keeps predicted inventories, IDs, intervals, and temporal matches fixed and deranges complete captions among predicted speakers. Mixtures with fewer than two predictions are excluded. The paired margin is original minus shuffled Binding; Appendix E.2 reports effective counts and standard errors.

The GT-assisted diagnostic adds reference speaker IDs and activity intervals to the audio and task prompt for the same 200 mixtures used in the free condition. Evaluation scores the generated outputs, including nonadherence and failures (Appendix E.3).

Conditional timeline quality T_{cond} computes frame-level Jaccard on the *Coverage* pairs using the same 10-ms grid and reference-boundary collars. It therefore differs from the main Timeline metric's separate assignment. Conditional statistics average first within a mixture, then over mixtures where they are defined. A mixture without accepted matches has an undefined conditional score and is excluded from that mean. Recall and precision use all mixtures in each group.

Conditional caption quality averages all 13 valid 1–10 scores per matched speaker, then averages speakers within a mixture and mixtures with at least one match. Incomplete judge responses must be completed before reporting these statistics. Conditional caption scores are not substituted into Multi or treated as an independent system ranking.

D.5 Judge configuration and prompts

Qwen3-235B scores one field at a time from the predicted text, human reference phrase, and field definition, without audio or system names. Reported decoding uses temperature 0 and a JSON response containing an integer score and a short rationale. All fields use semantic judging, including compact fields such as gender and age; there is no exact-tag scoring. The single-speaker track uses a 1–5 rubric, and the multi-speaker track uses 1–10 for all 13 fields, including personality.

Listings 1–4 reproduce verbatim the inference and judge prompts used in the reported runs. The surrounding explanatory text is not part of the model inputs. Personality is evaluated as a perceived vocal impression supported by the available evidence, not as verified identity or mental state.

Listing 1: Single-speaker captioning prompt used for the reported runs (part 1 of 2).

```
You are a Paralinguistics Analysis Expert trained to decode human vocal characteristics and
    speaking styles through acoustic patterns. Here is the instruction for your task.

### Instructions for Speech Style Caption

**Objective**: Based on the provided speech, generate a JSON-style output that includes
    concise descriptive phrase of the following vocal characteristcs. Please reply in English
    .

---

### Vocal Characteristics

- Gender: Speech characteristics associated with different gender identities, including vocal
    tract differences and socialized speech patterns.
- Pitch: The perceived frequency of sound, determining whether a voice sounds high or low. You
     can express relative pitch levels based on gender.
- Speed: How quickly or slowly someone is speaking. If the speaker exhibits specific rhythm
    patterns (such as "rap" "poem"), please indicate.
- Volume: How loudly or softly someone speaks. Examples range from soft to normal conversation
     volume to loud. Reserve "shouting" only for harsh or genuinely forceful vocalization.
- Age: An inference of the speaker's age group or life stage (such as infant, young child,
    child, teenager, young adult, middle-aged, elderly) based on vocal characteristics. For
    children, distinguish specific developmental stages: infant (babbling/very early speech),
     young child (developing speech, distinctly baby-like voice quality, unclear
    pronunciation, or slurred/connected speech patterns), child (clear speech patterns).
- Clarity: Whether pronunciation is distinct and precise or mumbled and blurred. Low clarity
    may involve slurring, mumbling, thick speech, with a swollen tongue, or running words
    together.
- Fluency: The smoothness and continuity of speech, indicating how naturally words flow
    without excessive hesitation, repetition, or filler words (such as "um," "uh," "like," "
    you know").
- Accent: Distinctive way of pronouncing words that indicates geographical origin,
    socioeconomic background, or non-native speaker status. If the accent is sufficiently
    pronounced, please specify the dialect region as precisely as you can. Otherwise response
     with the approximate region such as American English, British English or Mandarin
    Chinese.
- Texture: The timbral quality of a speaker's voice, including descriptors such as sweet,
    husky, deep, bright, warm, nasal, mellow, gravelly, or delicate. These attributes reflect
     both physiological traits and stylistic nuances, enabling differentiation between
    speakers or analysis of emotional/expressive tendencies.
- Emotion: The feeling expressed while speaking. Avoid exaggeration and overinterpretation.
    Most speech has mild to moderate emotional coloring, avoid extreme emotional terms (such
    as "desperate", "frantic") unless the emotion is genuinely overwhelming. Use measured
    terms like for typical emotional expression.
- Tone: The emotional or attitudinal quality conveyed through vocal inflection, encompassing
    patterns that signal nuances like sarcasm, formality, enthusiasm, or detachment.
- Personality: Please infer the speaker's overall personality based on the aforementioned
    voice characteristics, such as extroversion/introversion, confidence, assertiveness, or
    anxiety. Only describe the speaker's prominent and consistent personality traits evident
    in the speech.
- Summary: Identify the most notable characteristics of this audio or special overall style.
    If the speaker's vocal style clearly matches a recognizable character type or role (e.g.,
     TV host, robot, ghost, etc.), prioritize describing the character/role identity but no
    need to use the word "classic" or "typical". Avoid starting with "The speaker". Keep the
    summary concise and only focus on the special characteristics. Give your summary as a
    diverse descriptive sentence.

---
```

Listing 1 continued.

```
### Output Format

{
    "gender": "Gender of the speaker",
    "pitch": "Description of the speaker's pitch",
    "speed": "Description of the speaker's temporal flow",
    "volume": "Description of the speaker's volume"
    ......
}

---

### Final Checklist

- Make sure to note how the vocal characteristics change over time throughout the speech.
    Focus specifically on significant and unusual changes. Ignore minor or subtle variations.
- Ignore typical conversational emphasis or inflection patterns that normally occur unless
    they are extremely pronounced.
- The description should be concise phrase, avoid using single words or long sentences.
- When describing changes in characteristics, don't reference spoken content.
- Describe vocal characteristics using proportionate language. Reserve "dramatic" and "
    theatrical" only for speech that sounds like actual stage performance or deliberately
    exaggerated delivery. Animated vocal variety is not dramatic.
- Avoid exaggeration. Only point out features when they are clearly evident and distinctive.
- Please utilize diverse vocabulary and varied expressions, avoiding confinement to the
    illustrative examples provided.
- This speech is from film/TV content where performances are naturally more expressive than
    everyday conversation. Adjust descriptive standards accordingly, for example, only use "
    shouting" for genuinely aggressive vocalization, and "dramatic" only for truly heightened
     theatrical moments. Standard film/TV expressiveness should not be labeled as dramatic.
    Also, only use extreme emotional descriptors (desperate, frantic, ecstatic, devastated,
    furious, terrified) for genuinely overwhelming emotional displays.
```

Listing 2: Complete multi-speaker captioning prompt used for the reported runs (part 1 of 5).

```
You are a Paralinguistics Analysis Expert trained to decode human vocal characteristics and
     speaking styles through acoustic patterns. Here is the instruction for your task.

    ### Instructions for Multi-Speaker Speech Style Caption

    **Objective**: Based on the provided multi-speaker speech, generate one JSON-style output
     for each unique speaker. Each output item must include the speaker ID, all timestamp
     intervals where this speaker appears, and one overall speech-style caption for that
     speaker. Please reply in English.

    The `speech_caption` object should follow the same schema as the single-speaker speech
     caption format. Unlike segment-level captioning, do not generate a separate caption for
     every speech segment. Instead, merge all segments with the same `speaker_id` and generate
      only one `speech_caption` for that speaker.

    If a speaker has less than 0.5 seconds of total usable speech, set `"speech_caption": null
     `.

    ---

    ### Speaker Grouping Requirements

    - Group all speech segments by `speaker_id`.
    - There must be exactly one output item for each unique `speaker_id`.
    - Do not output repeated entries for the same speaker.
    - Preserve the original `speaker_id` exactly as provided.
    - Preserve all timestamp intervals for each speaker in `time_segments`.
    - Sort speakers by their first appearance in the audio.
    - Sort each speaker's `time_segments` by start time.
    - Generate one overall `speech_caption` based on all usable speech from that speaker.
    - For repeated segments from the same speaker, maintain consistency in stable vocal traits
      such as gender, age group, accent, and vocal texture.
    - For variable traits such as speed, volume, emotion, tone, clarity, and fluency, describe
      the speaker's overall tendency across their speech.
    - If the speaker's vocal characteristics change significantly across their own segments,
     mention the change concisely inside the relevant fields.
    - Do not describe the spoken content. Focus only on vocal characteristics and speaking
     style.
    - Do not infer one speaker's characteristics from another speaker.
    - If some individual segments are very short but the speaker has at least 0.5 seconds of
     total usable speech across all segments, still generate a speaker-level caption based on
     the usable acoustic evidence.
    - If all segments from a speaker are heavily overlapped, noisy, non-speech, or
     acoustically insufficient, set `"speech_caption": null`.
    - If a speaker only produces laughter, breathing, coughing, gasps, or very brief non-
     speech vocalizations, use `null` unless there is enough speech-like vocal evidence to
     support a reliable caption.

    ---
```

Listing 2 continued.

```
### Insufficient Speech Rule

    Use the total usable speech duration of each speaker after merging all segments with the same `
     speaker_id`.

    - If a speaker has less than 0.5 seconds of total usable speech, set `"speech_caption": null`.
    - If a speaker has 0.5 seconds or more of total usable speech, generate a speaker-level `
     speech_caption` unless the speech is heavily overlapped, noisy, non-speech, or otherwise
     acoustically insufficient.
    - Very brief backchannels, isolated words, laughter, breathing, coughing, gasps, or non-speech
     vocalizations should not count as sufficient speech unless they contain enough speech-like vocal
      evidence.
    - Do not infer a full speech caption from another speaker's voice.
    - When the available speech is brief but still above the 0.5-second threshold, use cautious and
     proportionate descriptions. Avoid over-specific claims about accent, personality, emotion, or
     age when the acoustic evidence is limited.
    - If a particular vocal attribute is not clearly supported by the audio, describe it
     conservatively instead of inventing details.

    ---
```

Listing 2 continued.

```
### Vocal Characteristics

    - Gender: Speech characteristics associated with different gender identities, including vocal
     tract differences and socialized speech patterns. Describe only the perceived vocal presentation
     , such as male-sounding, female-sounding, childlike, or ambiguous, rather than claiming the
     speaker's actual identity.
    - Pitch: The perceived frequency of sound, determining whether a voice sounds high or low. You
     can express relative pitch levels based on gender or age group.
    - Speed: How quickly or slowly someone is speaking. If the speaker exhibits specific rhythm
     patterns, such as "rap" or "poem", please indicate.
    - Volume: How loudly or softly someone speaks. Examples range from soft to normal conversation
     volume to loud. Reserve "shouting" only for harsh or genuinely forceful vocalization.
    - Age: An inference of the speaker's age group or life stage, such as infant, young child, child,
      teenager, young adult, middle-aged, or elderly, based on vocal characteristics. For children,
     distinguish specific developmental stages: infant, young child, and child.
    - Clarity: Whether pronunciation is distinct and precise or mumbled and blurred. Low clarity may
     involve slurring, mumbling, thick speech, swollen-tongue-like articulation, or words running
     together.
    - Fluency: The smoothness and continuity of speech, indicating how naturally words flow without
     excessive hesitation, repetition, or filler words, such as "um," "uh," "like," or "you know".
    - Accent: Distinctive way of pronouncing words that indicates geographical origin, socioeconomic
     background, or non-native speaker status. If the accent is sufficiently pronounced, specify the
     dialect region as precisely as possible. Otherwise, respond with the approximate language or
     region, such as American English, British English, or Mandarin Chinese. If the speech is too
     brief to support a precise accent judgment, use a cautious phrase such as "not clearly
     identifiable from the brief speech".
    - Texture: The timbral quality of a speaker's voice, including descriptors such as sweet, husky,
     deep, bright, warm, nasal, mellow, gravelly, delicate, breathy, crisp, or resonant.
    - Emotion: The feeling expressed while speaking. Avoid exaggeration and overinterpretation. Most
     speech has mild to moderate emotional coloring. Avoid extreme emotional terms, such as "
     desperate", "frantic", "ecstatic", "devastated", "furious", or "terrified", unless the emotion
     is genuinely overwhelming.
    - Tone: The emotional or attitudinal quality conveyed through vocal inflection, encompassing
     patterns that signal nuances like sarcasm, formality, enthusiasm, detachment, warmth, impatience
     , or authority.
    - Personality: Infer the speaker's overall personality based on the aforementioned voice
     characteristics, such as extroversion, introversion, confidence, assertiveness, anxiety,
     playfulness, calmness, or restraint. Only describe prominent and consistent personality traits
     evident in the speaker's voice across their speech. If the available speech is too brief to
     support a reliable personality inference, use a cautious phrase such as "limited evidence for
     personality inference".
    - Summary: Identify the most notable characteristics of this speaker's overall vocal style. If
     the speaker's vocal style clearly matches a recognizable character type or role, such as TV host
     , robot, narrator, ghost, child character, or announcer, prioritize describing the character or
     role identity. Do not use the word "classic" or "typical". Avoid starting with "The speaker".
     Keep the summary concise and focus only on distinctive vocal characteristics.

    ---
```

Listing 2 continued.

```
### Output Format

    Return a JSON array. Each item corresponds to one unique speaker.

        [
      {
        "speaker_id": "S0",
        "time_segments": [
          {
            "start_time": 0.00,
            "end_time": 3.42
          },
          {
            "start_time": 3.80,
            "end_time": 7.25
          }
        ],
        "speech_caption": {
          "gender": "Gender of the speaker",
          "pitch": "Description of the speaker's pitch",
          "speed": "Description of the speaker's temporal flow",
          "volume": "Description of the speaker's volume",
          "age": "Description of the speaker's age group",
          "clarity": "Description of pronunciation clarity",
          "fluency": "Description of speech fluency",
          "accent": "Description of accent or dialect",
          "texture": "Description of vocal timbre",
          "emotion": "Description of expressed emotion",
          "tone": "Description of vocal tone or attitude",
          "personality": "Description of prominent personality traits inferred from voice",
          "summary": "Concise overall description of the speaker's distinctive vocal style"
        }
      },
      {
        "speaker_id": "S1",
        "time_segments": [
          {
            "start_time": 3.42,
            "end_time": 3.80
          }
        ],
        "speech_caption": null
      }
    ]
```

Listing 2 continued.

```
### Final Checklist

    * Output one item per unique speaker, not one item per segment.
    * Merge all timestamp intervals belonging to the same speaker_id.
    * Keep time_segments complete and ordered.
    * Generate only one speech_caption for each speaker.
    * Use null when a speaker has less than 0.5 seconds of total usable speech.
    * Use null when the speaker's audio is non-speech, severely overlapped, extremely noisy, or
     acoustically insufficient.
    * If a speaker has at least 0.5 seconds of usable speech, generate a caption based on the
     available evidence, while keeping uncertain attributes cautious and non-specific.
    * Make sure to note how the speaker's vocal characteristics change across their own segments, but
      only when the change is significant and unusual.
    * Ignore typical conversational emphasis or inflection patterns unless they are extremely
     pronounced.
    * Keep each field as a concise phrase. Avoid single-word answers and avoid long sentences.
    * When describing changes in characteristics, do not reference spoken content.
    * Describe vocal characteristics using proportionate language.
    * Reserve \"dramatic\" and \"theatrical\" only for speech that sounds like actual stage
     performance or deliberately exaggerated delivery.
    * Reserve \"shouting\" only for genuinely aggressive or forceful vocalization.
    * Avoid exaggeration. Only point out features when they are clearly evident and distinctive.
    * Use diverse vocabulary and varied expressions. Do not rely only on the examples provided.
    * This speech may come from film or TV content where performances are naturally more expressive
     than everyday conversation. Adjust descriptive standards accordingly.
    * Output only the JSON array. Do not include explanations, markdown, or extra text.
```

Listing 3: Single-speaker judge prompt (1–5 rubric). The multi-speaker 1–10 template is reproduced in Listing 4.

```
[SYSTEM]
You are a meticulous, impartial evaluator of speech captions. You CANNOT
hear the audio. You are given a single attribute DIMENSION, its DEFINITION,
a REFERENCE description that is the verified ground truth for that
dimension, and a MODEL PREDICTION. Your job is to judge how well the
PREDICTION matches the REFERENCE for THIS dimension only.

Rules:
- Judge only the named dimension. Ignore information about other dimensions.
- The REFERENCE is correct by definition; the PREDICTION is what you score.
- Reward predictions that state the correct attribute value AND are at least
  as specific as the reference. Penalize contradictions, vagueness, hedging,
  wrong values, and hallucinated detail not supported by the reference.
- Treat clear synonyms / paraphrases as matches (e.g. "anxious" ~ "on edge",
  "raspy" ~ "gravelly"). Do not reward surface word overlap without semantic
  agreement.
- An empty, "unknown", off-topic, or contradictory prediction scores the
  minimum.
- Be terse. Output STRICT JSON and nothing else.

[USER]
DIMENSION: {dimension}
DEFINITION: {definition}
REFERENCE (ground truth): {reference}
MODEL PREDICTION: {prediction}

SCORING RUBRIC (integer 1-5):
5 = Fully consistent with the reference: correct value and equally or more
    specific; no contradictions.
4 = Correct main attribute with a minor omission, slightly less specificity,
    or a harmless extra detail.
3 = Partially correct: right coarse category but a salient qualifier is wrong
    or missing, or one correct cue mixed with one incorrect cue.
2 = Largely incorrect: only weak or generic overlap; the dominant attribute is
    wrong.
1 = Contradicts the reference, off-topic, empty, "unknown", or hallucinated.

First write a one-clause justification (<= 25 words), then the score.
Output ONLY this JSON object:
{"dimension": "{dimension}", "reasoning": "<short justification>", "score": <int 1-5>}

EXAMPLES
# good
DIMENSION: emotion | REFERENCE: worried, anxious | PREDICTION: the speaker
sounds anxious and on edge
-> {"dimension":"emotion","reasoning":"matches worried/anxious, equally specific","score":5}
# bad
DIMENSION: texture | REFERENCE: gravelly, deep | PREDICTION: a clear, normal
voice
-> {"dimension":"texture","reasoning":"contradicts gravelly/deep timbre","score":1}
```

Executed multi-speaker caption judge. Listing 4 reproduces the executed `caption_dim` template verbatim. Each request sends the rendered template as one user message, with no additional system message, using temperature 0. One request evaluates one of the 13 fields for one temporally matched speaker pair. Coverage and Timeline are computed deterministically; the caption judge does not select the speaker matches.

The placeholders are filled as follows: `DIMENSION` is the field name; `GT_TEXT` and `PRED_TEXT` are the canonicalized full reference and prediction speaker inventories in JSON; and `PAIR_CONTEXT` is a JSON object containing `match`, `gt_speaker`, and `pred_speaker`. The `match` entry contains the selected speaker IDs and temporal IoU; the speaker entries contain their intervals and captions. The `S0` IDs and score `1` in the output object illustrate the response format. The evaluation records each returned score against the requested speaker pair and field.

Listing 4: Executed multi-speaker caption judge prompt (1–10 rubric), with runtime placeholders retained.

```
You are a strict text-only judge for one dimension of one matched speaker pair.
You do not have audio. Compare only the GT speaker and Prediction speaker in PAIR_CONTEXT.
Do not infer new acoustic facts. Do not reward spoken-content descriptions.

Judge mode: caption_dim
Dimension: {{DIMENSION}}

Ground Truth and Prediction for full sample:
```text
{{GT_TEXT}}
```

```text
{{PRED_TEXT}}
```

Matched speaker pair:
```text
{{PAIR_CONTEXT}}
```

Only judge the dimension `{{DIMENSION}}`.

General score scale:
10 equivalent to GT, including important nuance.
9 nearly identical.
8 highly consistent with one minor omission or mild over-specificity.
7 mostly correct but misses a secondary detail.
6 basically correct category with noticeable deviation.
5 partially correct; captures one aspect but misses or distorts another.
4 weakly related; main description inaccurate.
3 mostly wrong.
2 contradictory or largely unrelated.
1 empty, impossible to judge, invented for null GT, or completely wrong.

Dimension guidance:
- gender: perceived vocal presentation only.
- age: life stage; penalize large age shifts.
- accent: language/dialect/accent family; penalize wrong language or region.
- pitch: low/mid/high and pitch movement.
- volume: soft/normal/loud/whisper/shouting/dynamic changes.
- speed: slow/moderate/fast, hesitation, rushing, deliberate pacing.
- texture: timbre such as bright, warm, breathy, nasal, husky, gravelly, deep, thin, resonant, crisp,
      strained.
- clarity: clear/articulate vs mumbled/slurred/blurred/noisy/overlapped.
- fluency: smooth/fluent vs hesitant/repeated/interrupted/filler-heavy.
- emotion: expressed emotion intensity and valence; penalize overdramatic unsupported emotion.
- tone: attitude such as authoritative, sarcastic, gentle, formal, impatient, soothing, questioning.
- personality: voice-inferred traits only; penalize unsupported personality overclaims.
- summary: concise preservation of the most distinctive GT vocal traits.

Output exactly one JSON object:
{
  "mode": "caption_dim",
  "dimension": "{{DIMENSION}}",
  "gt_speaker_id": "S0",
  "pred_speaker_id": "S0",
  "score": 1,
  "reason": "",
  "major_error": ""
}

`score` must be an integer from 1 to 10.
`major_error` should be empty if there is no major error; otherwise use one short label like
     wrong_value, unsupported_overclaim, missing_caption, wrong_null, or format_error.
No markdown. No extra text.
```

E JUDGE VALIDATION AND CONTROLLED DIAGNOSTICS

E.1 SCOPE OF HUMAN CALIBRATION

The reported calibration samples 50 items in each of seven open single-speaker fields: texture, clarity, fluency, emotion, tone, personality, and summary. Three annotators score these 350 field-specific items using the 1–5 rubric; the reported reference is their majority judgment. Judge–human Spearman correlation is $\rho = 0.79$ and mean absolute deviation is 0.40, compared with inter-annotator $\rho = 0.70$. Table S3 characterizes agreement on the reviewed source-caption annotations inherited by the multi-speaker references. No separate human study of mixture-level Binding was performed.

Matched paraphrases, unrelated reference descriptions, and contradictory values test whether scores move in the expected direction: respectively at least 4, at most 2, and the rubric minimum. The reported success rate is 88%. The secondary DeepSeek-V4-Pro judge uses the same script, prompt, and temperature 0. Its single-speaker scores show strong rank agreement with the Qwen judge (Spearman $\rho = 0.952$, Pearson 0.999). The multi-speaker secondary-judge evaluation covers 11 systems with identical prediction snapshots and complete secondary-judge fields. Gemini-3.8-Flash is excluded because its secondary run used a different prediction snapshot. The single-speaker comparison supports ranking-level robustness; field-level judge calibration remains dependent on the reported human sample.

Table S11: Secondary DeepSeek-V4-Pro judge on the single-speaker track. Scores are on the same 1–5 rubric; Norm is the 0–100 transformation.

System	Ovr.	Norm	n
Qwen3.5-Omni-Plus	3.771	69.3	1000
Qwen3.8-Omni-Flash	3.734	68.4	999
MOSS-Audio	3.643	66.1	992
Qwen3-Omni-Instruct	3.638	66.0	1000
Gemini-3.7-Flash	3.630	65.8	1000
Gemini-3.5-Flash	3.627	65.7	999
Gemini-3.8-Flash	3.622	65.6	1000
Gemini-3.1-Pro	3.612	65.3	1000
Qwen3-Omni-Thinking	3.585	64.6	1000
Gemini-3-Pro	3.508	62.7	999
Gemini-3.6-Flash	3.504	62.6	999
MiMo-V2.5	3.477	61.9	1000
Kimi-Audio	3.095	52.4	986
Voxtral-Small-24B	2.922	48.0	999

Table S12: Secondary DeepSeek-V4-Pro judge on the multi-speaker track for 11 systems evaluated on identical prediction snapshots. Coverage and Timeline are shared between judges; only Binding is rejudged.

System	Binding	Multi
Qwen3.5-Omni-Plus	0.511	70.3
Qwen3.8-Omni-Flash	0.500	68.2
Gemini-3.7-Flash	0.460	63.7
MiMo-V2.5	0.391	56.0
Gemini-3.5-Flash	0.394	54.2
Gemini-3.1-Pro	0.375	52.7
Qwen3-Omni-Thinking	0.373	51.9
MOSS-Audio	0.264	38.4
Qwen3-Omni-Instruct	0.235	32.6
Voxtral-Small-24B	0.125	19.3
Kimi-Audio	0.032	5.2
Gemini-3.8-Flash[†]	–	–

[†] Excluded because the secondary run used a different prediction snapshot.

E.2 CAPTION-SHUFFLE RESULTS

We use a stratified subset of 200 mixtures, stratified by speaker-count and the four overlap bins in Table S21. Each system retains its original speaker inventory, IDs, and activity intervals. Complete 13-field captions are randomly reassigned among predicted speakers with one no-fixed-point derangement per eligible mixture. Audio, references, matching rules, Coverage, and Timeline remain unchanged. Mixtures with fewer than two predicted speakers are excluded. We rerun the same Qwen3-235B judge and compute original minus shuffled Binding within each eligible mixture. Table S13 reports means and standard errors across eligible mixtures; the SEs do not account for dependencies from reused source clips.

Near-zero margins must still be interpreted alongside low baseline Binding, matching coverage, and the score floor. System-dependent eligible subsets preclude treating the margin as an independent capability ranking.

E.3 REFERENCE-STRUCTURE ASSISTANCE

Table 2 compares Qwen3.8-Omni-Flash on the same 200 mixtures. Providing reference speaker IDs and intervals raises Multi from 69.6 to 80.4. Table S14 records output adherence; all 200 mixtures enter each score, including failures. Matched-pair fields are fully scored in both conditions.

Table S13: Caption-shuffle results. n is the number of eligible mixtures. Printed means and margins are rounded separately.

System	n	Original B	Shuffled B	Margin	SE
Qwen3.8-Omni-Flash	198	0.507	0.302	+0.205	0.009
Qwen3.5-Omni-Plus	195	0.506	0.324	+0.183	0.008
Qwen3-Omni-Instruct	199	0.245	0.088	+0.157	0.010
Gemini-3.8-Flash	200	0.419	0.267	+0.153	0.008
Gemini-3.7-Flash	200	0.441	0.293	+0.149	0.009
Qwen3-Omni-Thinking	197	0.375	0.234	+0.141	0.008
MOSS-Audio	162	0.312	0.181	+0.131	0.013
Gemini-3.5-Flash	195	0.390	0.271	+0.119	0.008
MiMo-V2.5	183	0.403	0.294	+0.109	0.009
Gemini-3.1-Pro	196	0.379	0.282	+0.097	0.008
Voxtral-Small-24B	102	0.119	0.113	+0.007	0.003
Kimi-Audio	191	0.030	0.033	-0.003	0.002

Table S14: Output adherence in the free and GT-assisted conditions. Complete GT structure requires exact agreement of the generated speaker IDs and activity intervals with the reference.

Check	Free	GT-assisted
Evaluated mixtures	200	200
Nonempty recovered inventory	200	195
Complete GT structure reproduced	0	192
Missing matched-pair judge fields	0	0

The gain is a descriptive point estimate reported without a paired confidence interval. Assistance jointly supplies speaker-inventory and timing information and can improve Binding through matching coverage, so the gain does not isolate a timestamp-only effect. The free Multi score is 69.6 on these 200 diagnostic mixtures, compared with 67.5 on the full set of 1,000 mixtures. The caption-shuffle baseline Binding of 0.507 uses 198 eligible mixtures, whereas the free Binding of 0.503 in this diagnostic uses all 200.

E.4 Sensitivity to track aggregation

Across the 12 shared systems, single- and multi-speaker ranks have Spearman $\rho = 0.727$ and Kendall $\tau = 0.545$. Eight systems change rank, and 15 of the 66 system pairs reverse order.

Using the reported one-decimal track means, we vary the single-speaker weight in $\mathrm{VC}_w = w\tilde{S}+(1-w)\tilde{M}$ (Table S15). Qwen3.5-Omni-Plus, Qwen3.8-Omni-Flash, Gemini-3.7-Flash, and Gemini-3.8-Flash retain the first four positions at every setting. Geometric and harmonic track means reproduce the full default ordering (Table S19). These calculations leave the per-mixture Multi definition fixed.

Table S15: Ranking sensitivity to the single-speaker weight relative to the default $w = 0.4$. Scores are recomputed from the printed track means while keeping the per-mixture scoring protocol fixed.

Single-speaker weight w	0.3	0.4	0.5	0.6	0.7
Kendall τ vs. default	0.970	1.000	0.970	0.970	0.939
First four positions unchanged	Yes	Yes	Yes	Yes	Yes
Qwen3.5-Omni-Plus VC	70.5	70.7	71.0	71.2	71.5

Middle ranks depend on track weights. At $w = 0.6$, MiMo-V2.5 and Qwen3-Omni-Thinking differ by 0.08 using the printed means; their exact order requires unrounded inputs. Complete recalculations are provided as CSV alongside the manuscript.

F Complete Track Results and Difficulty Analysis

F.1 Single-speaker field scores

Tables S16 and S17 provide the exact printed values underlying Figure 3. Rows follow the overall score across 14 systems. Each field mean uses its available scored samples; the overall score gives equal weight to the 13 fields. Display rounding can cause a small difference between the average of printed field means and a printed aggregate. Parsing counts are in Table S9.

Table S16: Single-speaker scores (1–5). Perceived characteristics and prosodic fields.

System	Gender	Age	Accent	Pitch	Volume	Speed
Qwen3.5-Omni-Plus	4.748	4.158	4.556	3.848	3.781	3.873
Qwen3.8-Omni-Flash	4.629	4.082	4.354	3.819	3.832	3.880
Gemini-3.7-Flash	4.442	3.854	4.401	3.599	3.691	3.708
Qwen3-Omni-Instruct	4.691	3.939	4.519	3.672	3.736	3.679
Gemini-3.8-Flash	4.477	3.841	4.362	3.567	3.715	3.726
MOSS-Audio	4.630	3.905	4.566	3.704	3.751	3.670
Gemini-3.1-Pro	4.538	3.900	4.571	3.487	3.729	3.713
Gemini-3.5-Flash	4.535	3.868	4.470	3.619	3.806	3.570
Qwen3-Omni-Thinking	4.661	3.971	4.549	3.636	3.677	3.703
Gemini-3-Pro	4.262	3.795	4.412	3.528	3.573	3.581
Gemini-3.6-Flash	4.457	3.821	4.420	3.406	3.613	3.323
MiMo-V2.5	4.227	3.790	4.358	3.420	3.524	3.606
Kimi-Audio	4.455	3.485	3.058	3.301	3.343	3.211
Voxtral-Small-24B	3.595	3.392	3.905	2.841	3.046	3.029

Table S17: Single-speaker scores (1–5). Remaining fields; the summary is the thirteenth field.

System	Texture	Clarity	Fluency	Emotion	Tone	Person.	Summary
Qwen3.5-Omni-Plus	3.362	3.771	3.999	3.623	3.550	3.612	3.665
Qwen3.8-Omni-Flash	3.535	3.879	3.909	3.621	3.678	3.607	3.587
Gemini-3.7-Flash	3.429	3.795	3.964	3.510	3.533	3.556	3.436
Qwen3-Omni-Instruct	3.242	3.807	3.726	3.473	3.359	3.441	3.461
Gemini-3.8-Flash	3.372	3.769	3.875	3.498	3.488	3.569	3.435
MOSS-Audio	3.388	3.874	3.754	3.382	3.310	3.330	3.342
Gemini-3.1-Pro	3.240	3.837	3.810	3.477	3.481	3.419	3.391
Gemini-3.5-Flash	3.401	3.724	3.770	3.434	3.394	3.443	3.476
Qwen3-Omni-Thinking	3.275	3.674	3.625	3.321	3.293	3.251	3.344
Gemini-3-Pro	3.273	3.680	3.669	3.376	3.293	3.318	3.209
Gemini-3.6-Flash	3.210	3.589	3.662	3.276	3.318	3.301	3.325
MiMo-V2.5	3.101	3.751	3.766	3.311	3.300	3.322	3.169
Kimi-Audio	2.611	3.314	3.234	2.892	2.647	2.889	2.805
Voxtral-Small-24B	2.322	3.303	3.173	2.558	2.571	2.665	2.162

F.2 Multi-speaker components and unified scores

Table S18 gives all 12 systems and two full-span controls under the current protocol. The controls use empty captions, hence zero Binding and Multi. Table S19 combines the current track means; it also reports symmetric aggregations for the sensitivity analysis in Appendix E.4.

F.3 Single-speaker attribute slices

Table S20 reports eight overlapping slices for the four systems with available slice results. These systems rank 1, 4, 6, and 7 in the single-speaker comparison. Qwen3.5-Omni-Plus leads each reported slice, while the others change order. Rare accents are tags below 5% frequency in the single-speaker set. Other slices are defined by reference-phrase rules; the complete phrase-to-slice mappings are not reported. The "high emotion" label follows the reference-phrase rules; it does not denote a cal-

Table S18: Multi-speaker results on all 1,000 mixtures. C, T, B are component means; Multi averages per-mixture geometric scores. Recovery is the percentage with a nonempty parsed inventory. Confidence intervals use 2,000 mixture bootstrap resamples.

System	Multi	95% CI	C	T	B	Recovery (%)
Qwen3.5-Omni-Plus	69.7	$[68.6, 70.8]$	0.855	0.814	0.497	100.0
Qwen3.8-Omni-Flash	67.5	$[66.2, 68.7]$	0.829	0.782	0.484	99.5
Gemini-3.7-Flash	62.9	$[61.6, 64.1]$	0.782	0.737	0.442	100.0
Gemini-3.8-Flash	60.0	$[58.8, 61.3]$	0.746	0.707	0.423	100.0
MiMo-V2.5	55.5	$[53.9, 56.9]$	0.684	0.694	0.380	97.4
Gemini-3.5-Flash	53.8	$[52.4, 55.2]$	0.676	0.627	0.385	99.2
Gemini-3.1-Pro	52.3	$[50.7, 53.9]$	0.653	0.627	0.367	99.8
Qwen3-Omni-Thinking	51.4	$[49.8, 52.8]$	0.659	0.600	0.361	99.5
MOSS-Audio	38.5	$[36.8, 40.1]$	0.532	0.497	0.261	97.2
Qwen3-Omni-Instruct	32.5	$[31.3, 33.7]$	0.441	0.407	0.231	100.0
Voxtral-Small-24B	18.8	$[17.6, 20.0]$	0.307	0.296	0.117	100.0
Kimi-Audio	5.0	$[4.3, 5.7]$	0.074	0.192	0.028	99.9
Fixed-3 full-span	0.0	$[0.0, 0.0]$	0.226	0.374	0.000	100.0
GT-count full-span	0.0	$[0.0, 0.0]$	0.267	0.387	0.000	100.0

Table S19: Track aggregation for the 12 shared systems. VC $= 0.4\tilde{S} + 0.6\tilde{M}$, $G = \sqrt{\tilde{S}\tilde{M}}$, and $H = 2\tilde{S}\tilde{M}/(\tilde{S} + \tilde{M})$. All three are computed from unrounded track scores and rounded to one decimal; they produce the same ordering.

System	$\tilde{S}$	Multi	VC	G	H
Qwen3.5-Omni-Plus	72.2	69.7	70.7	71.0	70.9
Qwen3.8-Omni-Flash	71.9	67.5	69.3	69.7	69.7
Gemini-3.7-Flash	69.1	62.9	65.3	65.9	65.8
Gemini-3.8-Flash	68.6	60.0	63.5	64.2	64.1
MiMo-V2.5	64.7	55.5	59.2	59.9	59.7
Gemini-3.5-Flash	68.3	53.8	59.6	60.6	60.2
Gemini-3.1-Pro	68.4	52.3	58.7	59.8	59.3
Qwen3-Omni-Thinking	67.3	51.4	57.7	58.8	58.2
MOSS-Audio	68.5	38.5	50.5	51.3	49.3
Qwen3-Omni-Instruct	68.7	32.5	47.0	47.3	44.1
Voxtral-Small-24B	49.2	18.8	30.9	30.4	27.2
Kimi-Audio	54.3	5.0	24.7	16.5	9.2

ibrated intensity threshold. Slice scores are descriptive point estimates reported without confidence intervals.

Table S20: Single-speaker slice scores (1–5) for four evaluated systems. Cells average the 13 fields over clips in each slice. Slices overlap, and n denotes the total number of clips in each slice. Effective sample counts for individual systems and fields are not reported in this table. Qwen3.5+ = Qwen3.5-Omni-Plus; Qwen3-Ins. = Qwen3-Omni-Instruct; MOSS-A. = MOSS-Audio; Gem-3.1 = Gemini-3.1-Pro.

Slice	n	Qwen3.5+	Qwen3-Ins.	MOSS-A.	Gem-3.1
Rare accents	133	3.70	3.51	3.50	3.52
Child speech	75	3.73	3.63	3.60	3.50
Low clarity	350	3.64	3.52	3.51	3.46
High emotion	868	3.90	3.76	3.74	3.74
Dynamic pitch	436	3.89	3.72	3.70	3.73
Whispering	96	3.96	3.87	3.81	3.67
Shouting	44	3.67	3.45	3.34	3.51
Non-fluent	418	3.82	3.67	3.67	3.64

F.4 Multi-speaker difficulty slices

Table S21 reports Qwen3.5-Omni-Plus by speaker count and overlap. Recall and precision average over all mixtures; conditional columns use only mixtures with Coverage matches. The conditional caption column averages all 13 valid scores on matched speakers, whereas Binding also retains unmatched references with zero contribution; the two quantities therefore have different denominators and scales (Appendix D.4).

Recall decreases from 0.897 for two-speaker mixtures to 0.776 for five-speaker mixtures and rises to 0.877 for six-speaker mixtures. The six-speaker group contains 19 mixtures, all without overlap. Recall and conditional timeline quality also decrease across the overlap bins. These are descriptive comparisons between groups that differ in source and acoustic composition; conditional quality is evaluated only for matched speakers.

Table S21: Qwen3.5-Omni-Plus difficulty diagnostics. n_c/n denotes mixtures with defined conditional scores / all mixtures. T_c is conditional timeline Jaccard; Cap. is conditional caption quality (1–10). All 13 judge scores are required for every matched pair; missing judge responses are not assigned a floor score. No group confidence intervals are available.

Group	n_c/n	Recall	Precision	T_c	Cap.
2 speakers	389/394	0.897	0.904	0.921	6.362
3 speakers	305/306	0.837	0.873	0.908	6.263
4 speakers	163/164	0.782	0.855	0.896	6.357
5 speakers	116/117	0.776	0.825	0.906	6.157
6 speakers	19/19	0.877	0.875	0.891	6.191
No overlap	415/417	0.899	0.913	0.927	6.302
$0 < r \leq 0.1$	320/322	0.860	0.871	0.915	6.301
$0.1 < r \leq 0.2$	108/110	0.802	0.827	0.899	6.358
$r > 0.2$	149/151	0.698	0.825	0.862	6.275

G Comparison with Related Benchmarks

Table S22 compares related benchmarks by the outputs required from a system and the components evaluated by their scoring interfaces.

SpeechCraft, ParaSpeechCaps, and CapSpeech provide complementary speech–description resources for captioning or synthesis (Jin et al., 2024; Diwan et al., 2025; Wang et al., 2026). General audio benchmarks such as MMAU evaluate a broader range of audio understanding questions (Sakshi et al., 2025). VocalCap's contribution is the joint evaluation of vocal description and complete

Table S22: Related task interfaces and VocalCap's evaluation target. Selected questions, caption matching, and activity estimation provide evidence about different parts of the complete structured-output task.

Resource / protocol	Evaluated interface	Relation to VocalCap
MOSS-Audio (Yang et al., 2026a)	13-dimensional captions judged against human references	Closest field-wise description protocol; VocalCap adds evaluation of the complete multi-speaker structure.
SpeechCaps (Huang et al., 2025a)	Multi-talker caption supervision and generated answers about speaker attributes	Speaker–attribute associations through questions; VocalCap requires one caption and all intervals per speaker.
CompA (Ghosh et al., 2024)	Audio–caption matching for temporal composition and attribute binding	Tests selected relations; VocalCap additionally scores missing speakers and predicted activity intervals.
Diarization (Ryant et al., 2019; Fiscus et al., 2006)	Speaker activity over time	Temporal correspondence without the shared 13-field descriptive output.
VocalCap	Complete speaker inventory, activity intervals, and field-wise captions	Deterministic temporal matching, text-based field judging, and all-reference Binding.

speaker-grounded output, with separate coverage and timing diagnostics. Its controlled mixtures do not replace evaluation on natural conversations.

H Release, Reproducibility, and Responsible Use

Intended use and interpretation. VocalCap supports diagnostic evaluation of vocal description, speaker recovery, and temporal localization. Gender, age, accent, and personality refer to perceived presentation or impressions; they do not establish personal identity or psychological traits. Acted speech reflects performance choices rather than population-level attribute distributions. The benchmark should not be used to infer protected identity or mental state from an identifiable person's voice.

Annotation and reference provenance. Human-written reference phrases are retained for evaluation. LLM-assisted taxonomy induction maps these phrases to canonical tags for sampling and analysis without replacing them. Multi-speaker references inherit source descriptions, whose applicability after rendering is only partly audited (Appendix C). Text-only scores measure agreement with references, not independent acoustic verification (Appendix E).

Release scope and source rights. The release plan covers reference annotations, canonical tags, source pointers and timestamps, mixture composition metadata, complete API prompts and model strings, construction/evaluation code, and the aggregate result tables used in this paper. Raw film and television audio is not redistributed. Users must obtain source media independently and comply with the relevant access and use terms; source pointers do not convey media rights. Annotation and code licenses are stated separately from the source media's rights.

Available numerical evidence. The manuscript package contains the reported single-speaker field scores, multi-speaker components, caption-shuffle results, GT-assisted comparison, and available difficulty slices as machine-readable tables. Track-aggregation checks use printed rounded means and can be regenerated from these tables. Aggregate tables cannot substitute for per-sample predictions, judge responses, or audio-construction metadata when rerunning the benchmark.

Reproducibility record. The released configuration records the API model strings, prompts, decoding settings, mixture composition fields, diagnostic subset definitions, and judge outputs used for the reported tables. Aggregate tables are accompanied by machine-readable values and denominator conventions; the audio remains reconstructible only for users who hold the corresponding source media.