A Benchmark for Controllable Speaking-Style Captioning in Audio-Language Models

Shreeram Suresh Chandra[1,2*,3], Ashwyn Sharma[2], Liang Lu[2], Xiutian Zhao[1], Leda Sari[2], Björn Schuller [4], Berrak Sisman[1]

[1]Center for Language and Speech Processing, Johns Hopkins University, USA

[2]Otter AI [3]The University of Texas at Dallas, USA

[4]GLAM – Group on Language, Audio, & Music, Imperial College London, UK

[*]Work done during an internship at Otter AI.

Abstract

Speaking-style captioning (SSC) aims to generate natural language descriptions of *how* speech is delivered, capturing paralinguistic attributes such as vocal timbre, prosody, and expressivity. Many downstream applications, including conversational AI agents, controllable speech generation, and large-scale audio annotation, require dimension-specific style captions that describe targeted aspects of speech (e.g., speaker traits, emotion, or delivery style), rather than a single undifferentiated description. However, existing work lacks a unified formulation for controllable speaking-style captioning. We address this gap by formulating SSC as an instruction-following audio-language task, where explicit instructions specify the speaking-style dimensions to be described. We introduce *StyleInstructCaps*, the first standardized benchmark for controllable speaking-style captioning in audio-language models. *StyleInstructCaps* provides a dataset and evaluation framework that measures metadata groundedness, hallucination, instruction-following ability, generalization to unseen instructions and audio datasets, and speaker-style consistency. Across nine state-of-the-art audio-language models, we find that strong instruction-following ability does not necessarily imply grounded speaking-style understanding, and that SSC-specific instruction-diverse training substantially improves groundedness, controllability, and speaker consistency.[1]

1 Introduction

Speaking-style captioning (SSC) has recently emerged as a promising direction for generating natural-language descriptions of how speech is delivered, capturing properties such as vocal timbre, prosody, rhythm, emphasis, emotion, and other paralinguistic cues (Yamauchi et al., 2024). Despite growing interest, SSC still lacks a unified task formulation and evaluation paradigm.

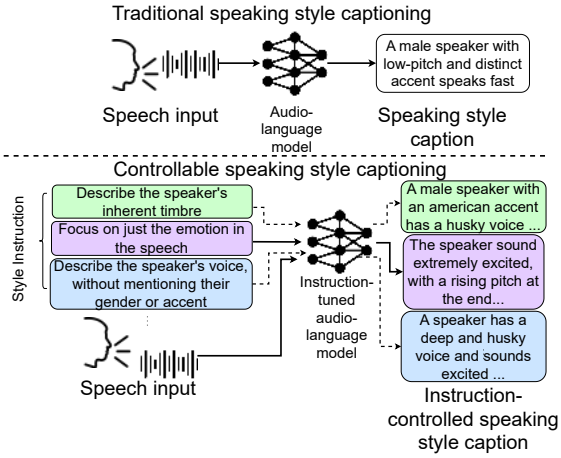


Figure 1: Speaking-style captioning as an instruction-following audio-language task

Researchers in audio understanding often treat SSC as an extension of speech comprehension, emphasizing communicative intent and discourse-level meaning (Jing et al., 2024; Chandra et al., 2025; Zhang et al., 2025b). In contrast, speech synthesis and generative modeling communities typically view speaking-style captions as controllable descriptions of vocal characteristics for expressive text-to-speech (Wang et al., 2025a; Guo et al., 2023) and voice conversion systems (Kuan et al., 2023; Hai et al., 2024). These differing perspectives reflect a deeper challenge: **speaking-style is inherently subjective** (Adigwe et al., 2024), spanning multiple interacting paralinguistic factors such as speaker identity, affect, prosody, and contextual delivery. As a result, existing SSC systems often capture different notions of style, making evaluation and comparison across models difficult.

To mitigate this ambiguity, we reformulate SSC as an **instruction-following audio-language task**, where explicit instructions specify which aspect of speaking-style the model should describe, as illustrated in Figure 1. Conventional SSC approaches

[1]Code, data, and benchmark are available at https://styleinstructcaps.github.io

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3820–3858

November 6-10, 2026 ©2026 Association for Computational Linguistics

typically produce a single holistic description of speech, often entangling speaker traits, emotion, prosody, and delivery characteristics, making captions difficult to interpret or evaluate systematically. In contrast, controllable SSC conditions generation on explicit style instructions, enabling the same utterance to yield multiple targeted and complementary descriptions depending on the task objective.

Despite their effectiveness on narrowly defined objectives, existing SSC models often adopt a single view of speaking-style, typically focusing on speaker-centric or emotion-centric attributes, which limits generalization to broader stylistic phenomena. Since speaking-style depends on downstream context, SSC models must be controllable and generate captions conditioned on explicit style instructions. Motivated by this perspective, we introduce *StyleInstructCaps*, the first standardized benchmark for controllable SSC. *StyleInstructCaps* evaluates generated captions along five dimensions: (1) **metadata groundedness**, (2) **hallucination**, (3) **instruction-following ability**, (4) **generalization to unseen speech data and prompts**, and (5) **speaker-style consistency**.

To support this benchmark, we construct a multi-style speaking-style caption corpus built on top of the ParaSpeechCaps (PSC) dataset (Diwan et al., 2025). The dataset contains 2.7k hours of speech, where each utterance is paired with six complementary caption types targeting distinct dimensions of speaking-style, including speaker identity, contextual delivery, emotion, pragmatics, listener perception, and holistic style summaries. These parallel annotations enable controlled evaluation of whether models can isolate specific stylistic dimensions, integrate multiple cues when required, and generalize across diverse speaking conditions.

Overall, this work makes three main contributions: (1) We present the first formulation of controllable SSC as an instruction-following audio-language task, in which style descriptions are generated conditioned on explicit stylistic instructions; (2) We introduce StyleInstructCaps, a standardized and challenging benchmark supported by a 2.7k-hour multi-style speaking-style caption corpus with six parallel caption types per utterance; (3) We evaluate nine state-of-the-art audio-language models, including Audio-Flamingo-3 (Goel et al., 2025), Voxtral (Liu et al., 2025), Qwen2-Audio (Chu et al., 2024), Qwen3-Omni (Xu et al., 2025), MOSS-Audio (Gong et al., 2026), Audio Flamingo Next (Ghosh et al., 2026), MERaLiON-AudioLLM (MERaLiON Team, 2025), and SALMONN (Tang et al., 2024), together with SSC-specific instruction-tuned variants of SALMONN. We find a consistent gap between instruction-following ability and grounded speaking-style understanding: models often generate stylistically plausible captions while hallucinating unsupported paralinguistic attributes, whereas SSC-specific, instruction-diverse training substantially improves groundedness, controllability, and speaker consistency.

Through this work, we aim to establish a unified foundation for controllable SSC and enable more faithful and interpretable speaking-style understanding in audio-language models.

2 Related Work

2.1 Speaking-Style Captioning for Paralinguistic Speech Understanding

Speaking-style captioning aims to generate natural-language descriptions of *how* speech is delivered, capturing attributes such as vocal timbre, prosody, emotion, and delivery style (Yamauchi et al., 2024). Unlike categorical paralinguistic tasks such as speaker recognition (Hansen and Hasan, 2015) and speech emotion recognition (Triantafyllopoulos et al., 2026), which assign a fixed label, and content-oriented tasks such as speech question answering (Peng et al., 2026) and speech summarization (Shang et al., 2024; Retkowski et al., 2025), which address what is said rather than how it is said, SSC produces open-ended descriptions that integrate multiple perceptual cues. This capability has become increasingly relevant in large audio-language models, where natural-language descriptions provide a flexible and interpretable interface for expressing subjective speaking-style phenomena (Su et al., 2026; Dhamyal et al., 2024; Lu et al., 2024, 2026). In the reverse direction, speech-generative audio-language models increasingly accept natural-language instructions to produce expressive speech, yet reliably realizing a requested emotion without degrading linguistic content remains challenging (Zhang et al., 2026; Zhao et al., 2026).

2.2 Datasets and Benchmarks

Progress in SSC depends on (1) speech corpora with rich speaking-style supervision and (2) evaluation protocols that assess generation quality beyond lexical overlap. Existing resources span multiple communities: expressive TTS and style-prompted

speech datasets provide speaker and delivery annotations, while paralinguistic benchmarks evaluate reasoning over cues such as emotion, prosody, and conversational pragmatics. Appendix A, Table 5 and Table 6 summarize representative datasets and benchmarks in this space. However, most existing benchmarks emphasize *recognition* rather than SSC as a standalone *generation* task, and do not support *controllable* evaluation over which stylistic dimensions should be described. This makes it difficult to compare SSC models consistently or evaluate whether models can selectively attend to requested style factors.

2.3 Speaking-Style Captioning Models

Advances in audio-language models enable speech-conditioned text generation (Cui et al., 2025), typically by coupling a speech encoder with a pre-trained LLM through an adapter (Verdini et al., 2025). SSC-relevant approaches fall into two families. (1) *General-purpose* audio-language models trained via multitask or instruction tuning (Das et al., 2025; Chu et al., 2023; Lu et al., 2026) can produce style-related descriptions (Huang et al., 2025), likely due to broad task coverage (e.g., ASR, AC, AST, SER, SV) (Tang et al., 2024). (2) *Task-specialized* SSC models explicitly target style caption generation: StyleCap (Yamauchi et al., 2024) combines a self-supervised speech encoder (Chen et al., 2022) with an auto-regressive LM; factor-conditioned methods introduce intermediate attribute prediction (Ando et al., 2024); emotion-focused SSC variants emphasize free-form generation and multimodal prompting (Xu et al., 2024a,b); and joint frameworks connect SSC with prompt-based speech synthesis (Zhu et al., 2024). Despite this progress, prior work rarely provides a standardized setting to test instruction-level controllability across multiple stylistic factors, motivating our benchmark design.

3 Methodology

We first formalize controllable SSC as an instruction-conditioned generation task (Section 3.1), then describe *StyleInstructCaps-DB*, the multi-dimensional caption dataset underlying the benchmark (Section 3.2). The evaluation tasks follow in Section 4.

3.1 Problem Formulation

We define *controllable SSC* as generating, for a given speech signal, a natural-language caption that describes only the speaking-style dimension requested by an explicit instruction (e.g., speaker traits, emotional expression, or delivery style). Unlike traditional SSC, which produces a single holistic description per utterance, controllable SSC requires models to selectively ground instruction-relevant attributes while avoiding unsupported or irrelevant ones.

Let $\mathcal{D} = \{(x_i, s_i, y_i)\}_{i=1}^{N}$ contain speech utterances x_i, free-form instructions s_i, and reference captions $y_i = (y_{i,1}, \ldots, y_{i,T_i})$. An audio-language model with parameters θ generates $\hat{y} \sim p_\theta(y \mid x, s)$. *StyleInstructCaps-DB* covers six speaking-style dimensions (Section 3.2), and the benchmark adds unseen instructions (Section 4).

3.2 Dataset Description

We introduce *StyleInstructCaps-DB*, a multi-style speaking-style caption dataset built on the existing SSC corpus ParaSpeechCaps (PSC) (Diwan et al., 2025). PSC aggregates multiple speech datasets and comprises two subsets: PSC-base, which includes Expresso (Nguyen et al., 2023), EARS (Richter et al., 2024), and VoxCeleb (Nagrani et al., 2017) and is human-annotated, and PSC-scaled, which consists of the Emilia dataset (He et al., 2024) and is automatically annotated. These annotations cover diverse intrinsic and situational speaking-style attributes across speakers and recording conditions. A brief overview of the PSC annotation procedure is provided in Appendix B.

While PSC provides a single caption for most utterances, *StyleInstructCaps-DB* introduces six complementary stylistic captions per utterance: (1) speaker idiosyncratic style, (2) situational or contextual delivery style, (3) expressive or emotional style, (4) linguistic-pragmatic style, (5) perceptual listener-centric style, and (6) an overall style summary. These caption types jointly span representative dimensions of speaking-style used in speech understanding and generation tasks.

Captions are generated using an LLM conditioned on the rich per-utterance metadata provided by PSC; an example metadata record is shown in the Appendix C Figure 4. *StyleInstructCaps-DB* is organized into train, validation, and evaluation splits. We use Mistral-Small-24B-Instruct (Mistral AI Team, 2025) to generate captions for the train and validation splits, and Llama-3.1-8B-Instruct (LlamaTeam, 2024) for the evaluation split, ensuring that benchmark evaluation does not benefit from exposure to captions generated by the same

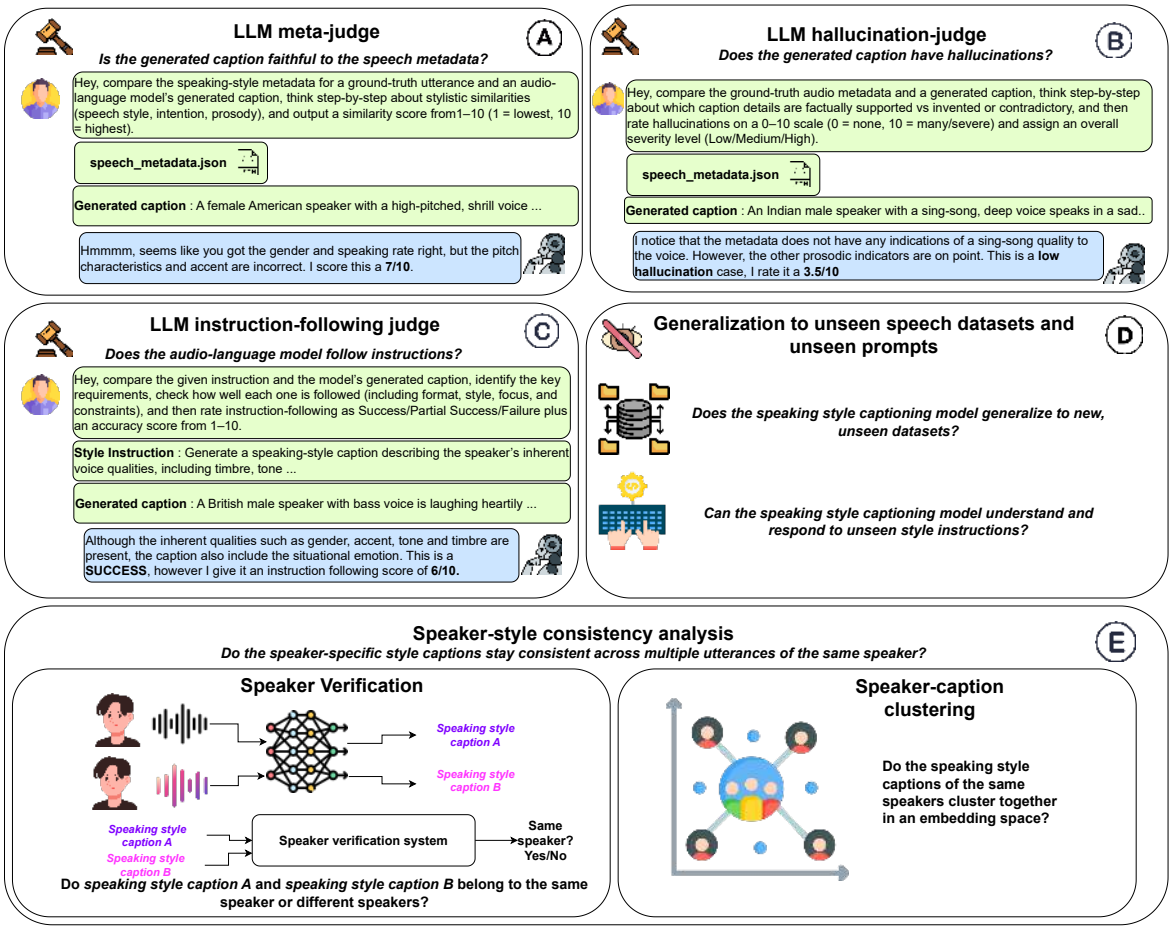


Figure 2: Overview of *StyleInstructCaps*, which evaluates models on metadata groundedness, hallucination, instruction-following, generalization to unseen prompts and speech data, and speaker-style consistency.

language model. Appendix C Figure 5 presents the captioning prompt, including explicit definitions of the six speaking-style dimensions, while Appendix C Table 7 reports instruction-caption pairs for the same utterance shown in Appendix C Figure 4.

4 StyleInstructCaps Benchmark

We introduce *StyleInstructCaps*, a benchmark built on *StyleInstructCaps-DB* for evaluating SSC models. The benchmark assesses captions along five evaluation dimensions (Figure 2): metadata groundedness, hallucination, instruction-following, generalization to unseen instructions and speech data, and speaker-style consistency. The following subsections describe the setup and evaluation for each dimension.

4.1 Metadata Groundedness

In the first evaluation task (Figure 2, Part A), we assess **metadata groundedness**, the extent to which a generated caption is supported by and consistent with the speech metadata. We sample utterances from the *StyleInstructCaps-DB* evaluation split with an emphasis on metadata richness, first selecting those with non-empty situational tags and then completing the set from the remaining pool. To ensure diversity, we cap each speaker at 10 utterances and limit any single accent to 25% of the set. We refer to this subset as *StyleInstructCaps-MetaSet* (distribution plots for accent, emotion, signal-to-noise ratio (SNR), and speech duration in Appendix C, Figure 6). Using the overall style summary instruction, an LLM judge panel compares the generated caption against PSC ground-truth metadata on speaker gender, accent, pitch, speaking rate, environment or noise level, and emotion or expression, with particular emphasis on prosodic alignment. The judge assigns a 1–10 similarity score with a brief justification, and we report the mean across utterances as the meta-score (prompt in Appendix H, Figure 9).

4.2 Hallucination

In the second evaluation task (Figure 2, Part B), we assess hallucination in generated speaking-style captions. We define hallucination as the inclusion of speaking-style attributes that are unsupported by

or contradictory to the input speech. Evaluation is conducted on *StyleInstructCaps-MetaSet*.

The LLM judge compares factual claims in the generated caption against the corresponding speech metadata, focusing on attributes such as speaker identity, emotion, accent, and acoustic characteristics. Descriptive or prosodic interpretations that do not contradict the metadata are treated leniently. The judge outputs a hallucination score on a 0 - 10 scale, together with a categorical severity label (Low, Medium, or High), reflecting both the frequency and severity of hallucinated attributes. The prompt is provided in Appendix H Figure 10.

4.3 Instruction-Following

We evaluate the controllability of SSC models through an explicit instruction-following task (See Figure 2 Part C). Using the *StyleInstructCaps-MetaSet* set, we assess whether models generate captions that adhere to user-specified style instructions. This evaluation probes instruction-following ability across the six speaking-style dimensions defined in *StyleInstructCaps-DB*, testing whether models can selectively attend to and describe the intended stylistic attributes. The LLM-based judge evaluates each generated caption for compliance with the given instruction, considering content relevance, stylistic focus, and constraint satisfaction. The judge outputs a categorical instruction-following outcome (Success or Failure) and a scalar score on a 1–10 scale, accompanied by a brief justification. We report the mean of this score as the *IFA score* (instruction-following accuracy, 1–10) and the percentage of captions judged a Success as the *IFA rate* (%). The complete evaluation prompts are provided in Appendix H Figure 11 and 12.

4.4 Generalization to Unseen Instructions

Beyond the six speaking-style dimensions in *StyleInstructCaps-MetaSet*, we assess generalization to instructions that differ *both in surface form and output requirements* from those seen during training. Training instructions are primarily narrative directives, whereas the unseen set is deliberately heterogeneous and includes short, deployment-oriented controls: slot-/tag-like schemas (e.g., YAML or Python-dict templates), strict JSON outputs, hard lexical and length constraints, numeric rating scales, and persona-conditioned prompts (e.g., call-center evaluator). This tests whether instruction tuning yields genuine instruction-following ability rather than overfitting to the verbose narrative style of the training distribution. We pair two utterances from each of 16 accent-diverse VCTK speakers (Yamagishi et al., 2019) with all 36 unseen instructions (Appendix H, Figure 13). Each of the resulting 1,152 outputs per system is scored with the instruction-following protocol and Qwen3-32B judge of Section 4.3, and we report the mean 1–10 score per system (Table 4).

4.5 Generalization to Unseen Speech Dataset

To evaluate SSC generalization to unseen speech datasets, we construct an evaluation set using the GigaSpeech corpus (Chen et al., 2021) with corresponding captions from Speechcraft (Jin et al., 2024). We sample 600 speech segments (300 from YouTube and 300 from podcasts), ensuring each segment originates from a distinct episode and is at least 3 seconds long. Each utterance is paired with a randomly sampled overall style summary speaking-style instruction, enabling evaluation under realistic acoustic conditions and diverse recording environments. Following the protocols in Sections 4.1 and 4.2, we assess metadata groundedness and hallucination with respect to the SpeechCraft captions. We note that Speechcraft provides limited metadata coverage compared to PSC.

4.6 Speaker-Style Consistency

Speaker-specific style captions describe inherent vocal traits and enable text-prompted speech synthesis via natural language control rather than explicit speaker IDs (Wang et al., 2025a).

To evaluate speaker-level consistency, we design two complementary evaluation schemes (See Figure 2 Part E). *Speaker-caption verification* measures pairwise consistency by testing whether captions generated from different utterances preserve invariant speaker-specific attributes. *Speaker-caption clustering analysis* evaluates the global structure of speaker-related information in the caption embedding space, examining whether captions from the same speaker form coherent and well-separated clusters.

Experiments are conducted on two datasets. The first is an in-domain balanced set (ID-Balanced) from the *StyleInstructCaps-DB* evaluation split, consisting of 400 utterances from speakers with at least 20 utterances each, with balanced gender and accent coverage (See Appendix C Figure 7 for distribution plots). The second is an unseen-speaker set (UD-VCTK) derived from VCTK, containing 320 utterances from 16 speakers spanning diverse

Table 1: Metadata groundedness, hallucination severity, and instruction-following performance evaluated on *StyleInstructCaps-MetaSet*. Blue shading indicates relative performance for scalar metrics. Metadata groundedness and hallucination are averaged over a panel of three LLM judges (Qwen3-32B, Gemma-4-31B-it, GPT-OSS-20B) with three runs per judge (9 runs per model); instruction-following is scored with a single Qwen3-32B run.

Model name	**LLM metadata groundedness**		**LLM hallucination severity (%)**			**Instruction following**	
	Meta score↑	**Hall.↓**	**Low**	**Mid**	**High**	**IFA score↑**	**IFA rate(%)↑**
Oracle Caption (PSC)	9.48	0.03	99.5	0.5	0	-	-
Oracle caption (*SIC-DB*)	9.86	0.22	99.8	0.2	0	8.48	71.3
Audio-Flamingo-3	3.88	6.43	15.4	20.7	63.8	8.21	62.4
Voxtral-Mini-3B	4.99	4.29	48.6	35.0	16.4	8.15	68.0
Voxtral-Small-24B	4.38	1.99	73.5	16.8	9.68	7.51	67.8
Qwen2-Audio-7B-Instruct	3.62	7.00	19.1	27.9	53.0	6.58	53.6
MERaLiON-AudioLLM	3.62	1.26	81.8	10.6	7.60	6.11	27.7
MOSS-Audio	6.32	5.50	21.8	45.4	32.7	8.82	82.8
Qwen3-Omni	5.75	3.86	51.2	30.6	18.2	7.99	74.0
Audio-Flamingo-next	5.09	5.54	19.35	38.94	41.7	7.59	48.5
SALMONN	4.68	4.66	42.1	28.3	29.5	7.50	58.5
SALMONN (PSC)	7.01	5.37	21.2	46.3	32.5	7.98	52.9
SALMONN (*SIC-DB*) - ST	7.30	4.33	35.2	37.3	27.4	8.37	65.6
SALMONN (*SIC-DB*) - MT	7.35	4.04	41.3	38.2	20.5	8.62	75.0

Closed-source reference — OpenAI-GPT-Audio: Meta score↑ 5.70, Hall.↓ 5.14.

Table 2: Speaker-style consistency evaluation using clustering and verification metrics on the ID-balanced and UD-VCTK sets. Darker shading indicates better performance per column. Higher is better for ARI, NMI, Purity, InterSep, and ROC-AUC; lower is better for EER.

Model	**ID-balanced**						**UD-VCTK**					
	ARI↑	NMI↑	Purity↑	InterSep↑	EER↓	AUC↑	ARI↑	NMI↑	Purity↑	InterSep↑	EER↓	AUC↑
Audio-Flamingo-3	0.009	0.046	0.100	0.046	0.476	0.532	0.011	0.048	0.121	0.057	0.463	0.556
Voxtral-Mini-3B	0.033	0.122	0.152	0.133	0.399	0.643	0.003	0.028	0.103	0.063	0.486	0.519
Voxtral-Small-24B	0.067	0.182	0.186	0.157	0.363	0.674	0.011	0.052	0.119	0.076	0.480	0.532
Qwen2-Audio-7B-Instruct	-0.002	0.010	0.072	0.026	0.498	0.499	-0.002	0.007	0.083	0.025	0.496	0.498
MERaLiON-AudioLLM	0.081	0.276	0.213	0.233	0.428	0.604	0.039	0.146	0.167	0.169	0.476	0.534
MOSS-Audio	0.191	0.381	0.347	0.221	0.282	0.792	0.076	0.228	0.186	0.175	0.319	0.755
Qwen3-Omni	0.152	0.297	0.243	0.204	0.376	0.630	0.069	0.216	0.184	0.165	0.311	0.734
Audio-Flamingo-Next	0.113	0.318	0.242	0.227	0.348	0.717	0.086	0.264	0.187	0.212	0.385	0.673
SALMONN	0.351	0.598	0.511	0.290	0.225	0.854	0.145	0.324	0.303	0.205	0.320	0.749
SALMONN (PSC)	0.297	0.541	0.491	0.271	0.231	0.861	0.157	0.340	0.312	0.209	0.320	0.711
SALMONN(*SIC-DB*) - ST	0.363	0.577	0.508	0.284	0.214	0.876	0.145	0.324	0.303	0.205	0.320	0.749
SALMONN(*SIC-DB*) - MT	0.357	0.590	0.517	0.279	0.212	0.876	0.156	0.341	0.298	0.222	0.282	0.795

accents. In both settings, each utterance is paired with the same 10 speaker-specific instructions applied to assess robustness to unseen speakers, accents, and recording conditions.

Speaker-Caption Verification. We formulate speaker-caption verification as a binary verification task in the text embedding space, analogous to speaker verification in the speech domain (Dehak et al., 2011). Positive trials pair captions from different utterances of the same speaker, while negative trials pair captions from different speakers. Captions are encoded using the pretrained E5-base-v2 text encoder (Wang et al., 2024), followed by attention-masked average pooling and L2 normalization. Speaker consistency is measured using cosine similarity between caption embeddings, with performance evaluated by Equal Error Rate (EER) and ROC-AUC.

Speaker-Caption Clustering Analysis. To assess global speaker organization, caption embeddings are clustered using k-means, with the number of clusters set equal to the number of speakers. Clustering quality is evaluated using adjusted rand index (ARI), normalized mutual information (NMI), and cluster purity. Inter-cluster separation is additionally reported to quantify global separation between speakers.

4.7 Are LLM judges reliable for the SSC task evaluation?

For metadata groundedness, hallucination, instruction-following, and generalization, we adopt the LLM-as-a-judge paradigm (Zheng et al., 2023), which prior work has shown to align better with human preferences than automatic metrics such as ROUGE (Lin, 2004), CIDEr (Vedantam

et al., 2015), BERTScore (Zhang* et al., 2020), and n-gram based measurements (Mondshine et al., 2025). We validate this choice with three experiments.

Adversarial gaming check. Following the LLM judge reliability protocol of EmoSura (Jing et al., 2026), we introduce controlled poisoning perturbations into the oracle captions. The resulting score shifts show that the judges detect such degradation in caption quality (Appendix D).

Panel of LLM judges. To ensure that the reported scores and system rankings are not an artifact of a single evaluator, we use the panel of three judges described in Section 5. All judge pairs rank the systems consistently (Spearman $\rho \geq 0.864$), so the metadata-groundedness and hallucination scores in Tables 1 and 3 are averaged over the full panel (Appendix F).

Human agreement with the judge. In a pairwise preference study, thirteen listeners made forced choices between captions of different SSC systems over 100 caption pairs spanning all 14 systems in Table 1, while hearing only the raw audio. The human majority vote agrees with the judge panel on 85.9% of pairs (Cohen's $\kappa = 0.717$), rising from 73.3% on near-ties to 93.3% on the largest score gaps, and a Bradley–Terry ranking fit to the human votes recovers the judge system ordering at Spearman $\rho = 0.895$. The protocol and full results are given in Appendix E.

5 Experiment Setup

We use a panel of three open-weight LLM judges from different model families: Qwen3-32B (Yang et al., 2025) (thinking), Gemma-4-31B-it (thinking) (Team et al., 2026), and GPT-OSS-20B (reasoning) (OpenAI, 2025). Open-weight judges enable reproducible, open evaluation compared to closed-source alternatives. Metadata groundedness and hallucination are averaged over this panel, with three runs per judge on *StyleInstructCaps-MetaSet* and two runs per judge on the out-of-domain evaluation sets (GigaSpeech, VCTK). Instruction-following is scored with a single Qwen3-32B run because of its high computational cost of 2,604 judgments per audio-language model, judge, and run (Appendix F).

We evaluate nine open-source audio-language models: Audio-Flamingo-3, Voxtral-Mini-3B, Voxtral-Small-24B, Qwen2-Audio-7B-Instruct, MERaLiON-AudioLLM, Audio-Flamingo-Next, MOSS-Audio, Qwen3-omni and SALMONN (specifications in Appendix I, Table 19). As a closed-source reference, we additionally evaluate OpenAI-GPT-Audio; we report only its core groundedness metrics, as the fine-grained analyses require large-scale inference with a paid API. We further instruction-tune SALMONN under three configurations: (1) using the original ParaSpeechCaps captions, SALMONN (PSC); (2) using the overall style summary captions from *StyleInstructCaps-DB* in a single-task setup, SALMONN (*SIC-DB*)-ST; and (3) using the multi-caption set from SIC-DB with multiple instruction-caption pairs per utterance in a multi-task setup, SALMONN (*SIC-DB*)-MT. All variants are trained with supervised fine-tuning (SFT), minimizing the token-level negative log-likelihood over their respective training set $\mathcal{D}$ (Section 3.1):

$$\mathcal{L}_{\text{SFT}}(\theta) = \mathbb{E}_{(x,s,y)\sim\mathcal{D}} \left[\sum_{t=1}^{T} -\log p_\theta(y_t \mid x, s, y_{<t}) \right].$$

Additional training and inference details are in Appendix I. Data pre-processing follows the ParaSpeechCaps setup (Diwan et al., 2025), and split statistics are in Appendix C, Table 8. We further include an analysis of lexical diversity of the generated captions in the Appendix J, reporting HD-D, Self-BLEU, Distinct-2, Distinct-3, and caption length statistics (mean and standard deviation) - to measure the stylistic richness of the speaking style captions.

6 Results and Discussion

Table 1 reports the results for the first three evaluation tasks – metadata groundedness, hallucination severity and instruction-following, all evaluated by LLM judges. Broadly, we observe that **general-purpose SOTA large audio-language models tend to struggle with the SSC task**, even though they have been trained in a large-scale setting with iterative training strategies. Metadata groundedness and hallucination severity are largely independent across models (Spearman $\rho \approx 0$): some weakly grounded models hallucinate rarely (e.g., MERaLiON-AudioLLM, Voxtral-Small-24B) while others hallucinate heavily (e.g., Qwen2-Audio, Audio-Flamingo-3), so a low hallucination score alone does not indicate faithful style description. SSC-specific training

yields the largest gains in groundedness: all three SSC-trained SALMONN variants exceed the best general-purpose model, MOSS-Audio. Appendix K Figure 14 provides an example of judgment for the metadata groundedness evaluation. Pairwise significance tests over utterances are reported in (Appendix G).

Interestingly, models trained with SSC-relevant audio encoders and tasks outperform those exposed only to ASR, which captures *what* is said but not *how* it is said. For instance, Audio-Flamingo-3 achieves state-of-the-art results on MMAU (Sakshi et al., 2025) and MMAR (Ma et al., 2025) but shows a low metascore and high hallucination on SSC, while SALMONN, despite a smaller backbone, benefits from its BEATs (Chen et al., 2023) + Whisper (Radford et al., 2023) encoder and SSC-relevant training tasks (SQA, GR, SER).

Next, we examine instruction-following performance. As shown in Table 1, instruction-following and metadata groundedness are moderately correlated across models,[2] but the relation is asymmetric: well-grounded models also follow instructions well, whereas strong instruction-following does not guarantee grounding. Hallucination severity is largely unrelated to instruction-following. Performance is primarily driven by the diversity of tasks seen during instruction tuning, the capacity of the underlying LLM backbone, and hyperparameter choices that prevent overfitting to narrow instruction distributions. MOSS-Audio is the strongest general-purpose model on both axes, showing that the two abilities can co-occur. In contrast, Audio-Flamingo-3 (IFA score 8.21) and Voxtral-Mini-3B (8.15) achieve strong instruction-following despite weak faithfulness-oriented performance (meta scores of 3.88 and 4.99). **This dissociation indicates that generic instruction tuning suffices to produce stylistically well-formed captions but fails to ground them in speech-specific paralinguistic attributes**. SSC-specific training alone does not improve instruction-following (IFA rate of 52.9% for SALMONN (PSC) vs. 58.5% for SALMONN), whereas instruction-diverse training does: SALMONN (*SIC-DB*)-MT raises the IFA score from 7.50 to 8.62 and the IFA rate from 58.5% to 75.0%.

On the unseen speech dataset (Table 3), SSC-specific training brings no further groundedness gain: all SALMONN variants score 6.25–6.30, on par with base SALMONN (6.28; not significant, Appendix G), while MOSS-Audio scores highest (6.70). Pooled instruction-following scores also hide a large effect of output format. We label each of the 36 unseen instructions by required format (narrative, key-value slots, tags and lists, minimal imperatives, persona) and re-score every system with the same Qwen3-32B judge. Table 4 shows that format sensitivity is model-specific: Audio-Flamingo-3 and Qwen2-Audio lose much of their controllability on key-value slots (3.34 and 4.00 vs. 7.26 and 7.27 on narrative), the Voxtral models drop on minimal imperatives (4.85 and 3.46), and MOSS-Audio and Qwen3-Omni remain strong across all formats. All SALMONN models, including the untuned baseline, score higher on key-value and tag/list schemas than on narrative instructions, so this pattern stems from the backbone rather than SSC-specific training.

Table 3: Evaluation on the unseen speech dataset. Metadata groundedness (1–10, ↑) and hallucination count (0–10, ↓) are averaged over the three-judge panel (Qwen3-32B, Gemma-4-31B-it, GPT-OSS-20B) with two runs per judge (6 runs per model) on the GigaSpeech-derived out-of-domain evaluation set.

Model	Meta score↑	Hall.↓
OpenAI-GPT-Audio	4.98	7.02
Audio-Flamingo-3	3.85	5.97
Voxtral-Mini-3B	4.57	7.12
Voxtral-Small-24B	3.45	6.45
Qwen2-Audio-7B-Instruct	4.62	6.97
MERaLiON-AudioLLM	3.95	5.72
MOSS-Audio	6.70	5.90
Qwen3-Omni	4.40	3.81
Audio-flamingo-next	5.81	6.65
SALMONN	6.28	6.29
SALMONN (PSC)	6.28	6.38
SALMONN (*SIC-DB*) - ST	6.25	5.93
SALMONN (*SIC-DB*) - MT	6.30	6.05

Next, we turn to discussing speaker style consistency. Table 2 reveals clear and consistent trends in speaker-style consistency across both the ID-balanced set and UD-VCTK sets. **SALMONN-based models yield the most speaker-consistent captions**, and this advantage is already present in the untuned baseline. All SALMONN variants outperform all other models on every metric on the ID-balanced set and on ARI, NMI, and Purity on UD-VCTK. On UD-VCTK, however, Qwen3-Omni

[2]Spearman $\rho = 0.68$ over the nine general-purpose models ($\rho = 0.73$ over all twelve systems); between hallucination severity and instruction-following, $\rho = 0.27$. Given the small number of systems, we report these as descriptive statistics.

Table 4: Instruction-following on unseen prompts, broken down by the output format each instruction requires. Each cell aggregates 32 clips over the instructions in that family, giving 128 to 416 judgements per system and 1,152 per system in total.

Model	Narrative	Key-value	Tags/list	Min. imper.	Persona
Audio-Flamingo-3	7.26	3.34	4.12	6.39	7.47
Voxtral-Mini-3B	6.23	6.30	7.31	4.85	7.41
Voxtral-Small-24B	5.43	7.68	7.30	3.46	5.82
Qwen2-Audio-7B-Instruct	7.27	4.00	5.63	5.84	6.96
MERaLiON-AudioLLM	4.50	6.91	5.70	5.15	5.79
MOSS-Audio	8.00	8.57	8.34	8.39	8.33
Qwen3-Omni	7.08	8.89	9.07	7.89	7.72
Audio-flamingo-next	6.69	7.62	7.52	6.38	7.06
SALMONN	6.30	7.86	7.45	6.90	7.67
SALMONN (PSC)	6.57	7.77	7.52	5.35	7.14
SALMONN (*SIC-DB*) - MT	6.69	7.85	8.28	6.67	7.68

and MOSS-Audio achieve lower EER than every variant except SALMONN (*SIC-DB*)-MT, which is best on EER (0.282) and AUC (0.795), and Audio-Flamingo-Next exceeds all variants but MT on InterSep. These results indicate that SALMONN-based captions encode stable, speaker-specific paralinguistic attributes rather than surface-level descriptions. The t-SNE (Maaten and Hinton, 2008) visualizations of speaker-caption embeddings for a subset of six speakers from the UD-VCTK set (Figure 3) show more coherent and well-separated speaker clusters for SALMONN-based models than for the other models. A full t-SNE analysis across all speakers is provided in Appendix L Figure 15.

Relative to the untuned baseline, style-caption fine-tuning yields mixed, modest changes: SALMONN (PSC) improves clustering on UD-VCTK but not on the ID-balanced set, while the *SIC-DB* variants improve verification (EER, AUC) on the ID-balanced set. Single-task instruction tuning (ST) further improves consistency on the ID-balanced set but does not generalize to UD-VCTK, indicating over-specialization to a narrow instruction distribution. In contrast, **multi-task instruction tuning (MT) yields the most balanced and transferable performance**, with strong clustering structure and competitive verification scores across both datasets. Finally, the strong agreement between clustering and verification metrics indicates that these improvements reflect a well-structured caption embedding space. **Overall, speaker consistency stems mainly from the SALMONN backbone; among fine-tuning strategies, instruction-diverse (MT) training transfers best across domains.**

7 Conclusions

We introduced *StyleInstructCaps*, a benchmark that formulates controllable SSC as an instruction-following audio-language task. Its dataset, *StyleInstructCaps-DB*, extends ParaSpeechCaps with six complementary captions per utterance, covering speaker idiosyncratic, situational, expressive, linguistic-pragmatic, perceptual, and overall style. The benchmark evaluates models along five axes: metadata groundedness, hallucination, instruction-following, generalization to unseen instructions and speech datasets, and speaker-style consistency. For groundedness and hallucination, we designed an LLM-as-a-judge protocol that checks captions against per-utterance speech metadata. We validated this protocol with adversarial perturbation checks, consistent system rankings across a three-judge panel (Spearman $\rho \geq 0.864$), and a listening study in which the human majority vote agreed with the panel on 85.9% of caption pairs. We measured speaker-style consistency directly in the caption embedding space, using speaker-caption verification and clustering to test whether captions of the same speaker preserve invariant vocal traits across utterances. Together, these components turn SSC from a loosely defined capability into a measurable research task.

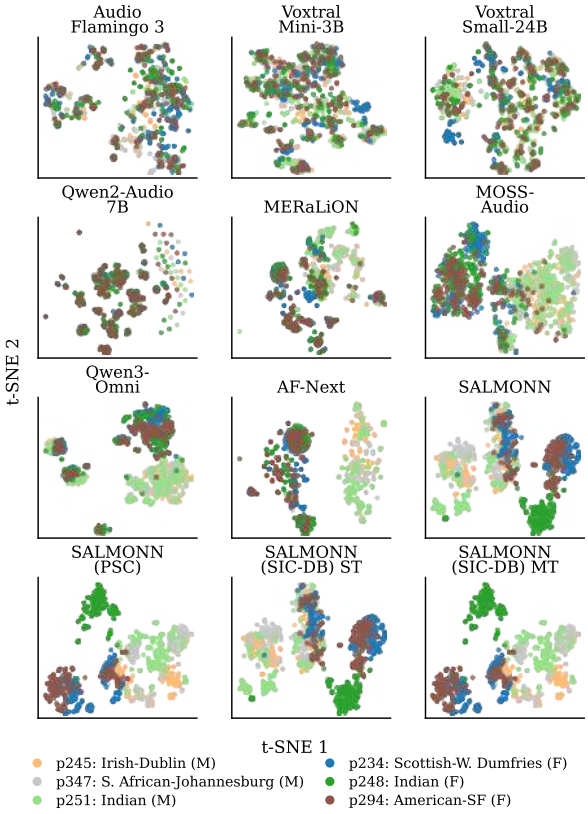


Figure 3: t-SNE visualization of speaker-style caption embeddings on UD-VCTK (6 speaker subset)

Limitations

This study is subject to several limitations inherent to large-scale speaking-style captioning (SSC). First, publicly available speech corpora with rich, fine-grained, per-utterance speaking-style annotations remain limited in both scale and diversity. As a result, SSC systems are often trained and evaluated on datasets with constrained stylistic coverage, which may not fully capture the variability of real-world expressive speech across speakers, domains, recording conditions, and cultural contexts.

Second, the proposed evaluation framework relies substantially on LLM-based judgment for assessing open-ended caption quality, including groundedness, hallucination, and instruction-following behavior. Although recent LLM judges exhibit strong reasoning and language understanding capabilities, their evaluations remain dependent on the specific judge model, prompting strategy, and the distribution of examples encountered during pretraining and instruction tuning. Consequently, evaluation outcomes may vary across alternative judge architectures or future generations of language models. To reduce this dependence, we average the open-ended metrics over a panel of three judges from different model families, whose system rankings agree closely (pairwise Spearman ρ of 0.864–0.978; Appendix F).

Third, despite efforts to improve robustness through neutral prompt design, prompt paraphrasing, and multi-prompt evaluation, LLM-based assessments may still exhibit sensitivity to prompt wording, decoding configurations, or stylistic biases in generated text. While our adversarial calibration experiments demonstrate that the judge is sensitive to targeted perturbations, such analyses cannot fully eliminate all forms of evaluation instability or hidden biases.

Fourth, the overall pipeline introduces a potential closed-loop evaluation concern, since metadata-derived descriptions are used both to construct reference captions and to evaluate generated outputs through LLM-based scoring. Although we explicitly validate the sensitivity of the judge to adversarial manipulations and metadata inconsistencies, some degree of implicit alignment between the generation and evaluation processes may still remain. Future work should further investigate evaluation paradigms involving independent human annotation and audio-grounded verification.

Finally, the current benchmark primarily evaluates metadata-grounded captioning rather than fully audio-grounded speaking-style understanding. In many cases, the reference captions reflect structured metadata annotations associated with the utterance rather than exhaustively capturing all perceptually salient properties present in the speech signal itself. As a result, models may partially optimize toward reproducing metadata-aligned descriptions instead of developing deeper holistic representations of expressive speech. Bridging this gap between metadata supervision and truly audio-grounded style understanding remains an important direction for future SSC research.

Ethical Considerations

In this work, we utilize open-source audio-language models to study the task of speaking-style captioning. We emphasize that the paralinguistic understanding capabilities of large language models remain an open research problem and are not yet sufficiently reliable for high-stakes use. Consequently, the outputs and evaluations presented in this paper should not be interpreted as substitutes for human judgment in scenarios that require factual accuracy, mental-state inference, or sensitive assessments of speaker intent or affect. Our results are intended to support research and benchmarking in controlled experimental settings, rather than deployment in applications where misinterpretation of paralinguistic cues could lead to harm.

References

Adaeze Adigwe, Sarenne Wallbridge, and Simon King. 2024. What do people hear? Listeners' Perception of Conversational Speech. In *Interspeech 2024*, pages 1210–1214.

Atsushi Ando, Takafumi Moriya, Shota Horiguchi, and Ryo Masumura. 2024. Factor-Conditioned Speaking-Style Captioning. In *Interspeech 2024*, pages 782–786.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. *Advances in Neural Information Processing Systems*, 37:56898–56918.

Shreeram Suresh Chandra, Lucas Goncalves, Junchen Lu, Carlos Busso, and Berrak Sisman. 2025. EmotionRankCLAP: Bridging Natural Language Speaking Styles and Ordinal Speech Emotion via Rank-N-Contrast. In *Interspeech 2025*, pages 3000–3004.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*, pages 3670–3674.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5178–5193. PMLR.

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. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

Wenqian Cui, Xiaoqi Jiao, Ziqiao Meng, and Irwin King. 2025. Voxeval: Benchmarking the knowledge understanding capabilities of end-to-end spoken language models. *Preprint*, arXiv:2501.04962.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, Zhaocheng Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, Xilai Li, Karel Mundnich, Monica Sunkara, Sravan Bodapati, Sundararajan Srinivasan, Kyu J Han, and Katrin Kirchhoff. 2025. Speechverse: A large-scale generalizable audio language model. *Preprint*, arXiv:2405.08295.

Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2011. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Yayue Deng, Guoqiang Hu, Haiyang Sun, Xiangyu Zhang, Haoyang Zhang, Fei Tian, Xuerui Yang, Gang Yu, and Eng Siong Chng. 2025. Multi-bench: A multi-turn interactive benchmark for assessing emotional intelligence ability of spoken dialogue models. *Preprint*, arXiv:2511.00850.

Hira Dhamyal, Benjamin Elizalde, Soham Deshmukh, Huaming Wang, Bhiksha Raj, and Rita Singh. 2024. Prompting audios using acoustic properties for emotion representation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11936–11940. IEEE.

Anuj Diwan, Zhisheng Zheng, David Harwath, and Eunsol Choi. 2025. Scaling rich style-prompted text-to-speech datasets. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3639–3659, Suzhou, China. Association for Computational Linguistics.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang-gil Lee, Jaehyeon Kim, Aya Aljafari, and 1 others. 2026. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*, arXiv:2604.10905.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Preprint*, arXiv:2507.08128.

Yitian Gong, Kuangwei Chen, Zhaoye Fei, Xiaogui Yang, Ke Chen, Yang Wang, Kexin Huang, Mingshu Chen, Ruixiao Li, Qingyuan Cheng, and 1 others. 2026. Moss-audio-tokenizer: Scaling audio tokenizers for future audio foundation models. *arXiv preprint arXiv:2602.10934*, arXiv:2602.10934.

Wenhao Guan, Yishuang Li, Tao Li, Hukai Huang, Feng Wang, Jiayan Lin, Lingyan Huang, Lin Li, and Qingyang Hong. 2024. Mm-tts: Multi-modal prompt based style transfer for expressive text-to-speech synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18117–18125.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jiarui Hai, Karan Thakkar, Helin Wang, Zengyi Qin, and Mounya Elhilali. 2024. DreamVoice: Text-Guided Voice Conversion. In *Interspeech 2024*, pages 4373–4377.

John HL Hansen and Taufiq Hasan. 2015. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal processing magazine*, 32(6):74–99.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, and 1 others. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE.

Chien-Yu Huang, Min-Han Shih, Ke-Han Lu, Chi-Yuan Hsiao, and Hung-Yi Lee. 2025. Speechcaps: Advancing instruction-based universal speech models with multi-talker speaking style captioning. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2024. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10301–10305. IEEE.

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

Xin Jing, Andreas Triantafyllopoulos, and Björn Schuller. 2024. ParaCLAP – Towards a general language-audio model for computational paralinguistic tasks. In *Interspeech 2024*, pages 1155–1159.

Xin Jing, Andreas Triantafyllopoulos, Jiadong Wang, Shahin Amiriparian, Jun Luo, and Björn Schuller. 2026. Emosura: Towards accurate evaluation of detailed and long-context emotional speech captions. *Preprint*, arXiv:2603.09820.

Masaya Kawamura, Ryuichi Yamamoto, Yuma Shirahata, Takuya Hasumi, and Kentaro Tachibana. 2024. LibriTTS-P: A Corpus with Speaking Style and Speaker Identity Prompts for Text-to-Speech and Style Captioning. In *Interspeech 2024*, pages 1850–1854.

Chun-Yi Kuan, Chen-An Li, Tsu-Yuan Hsu, Tse-Yang Lin, Ho-Lam Chung, Kai-Wei Chang, Shuo-Yiin Chang, and Hung-yi Lee. 2023. Towards general-purpose text-instruction-guided voice conversion. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, and 87 others. 2025. Voxtral. *Preprint*, arXiv:2507.13264.

AIMeta LlamaTeam. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, He Huang, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung yi Lee. 2024. DeSTA: Enhancing Speech Language Models through Descriptive Speech-Text Alignment. In *Interspeech 2024*, pages 4159–4163.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, Yi-Cheng Lin, Yu-Xiang Lin, Chi-An Fu, Chun-Yi Kuan, Wenze Ren, Xuanjun Chen, Wei-Ping Huang, En-Pei Hu, Tzu-Quan Lin, and 9 others. 2026. Desta2.5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *IEEE Transactions on Audio, Speech and Language Processing*, 34:2062–2076.

Dan Lyth and Simon King. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *Preprint*, arXiv:2402.01912.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, and 15 others. 2025. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Gallil Maimon, Amit Roth, and Yossi Adi. 2025. Salmon: A suite for acoustic language model evaluation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

MERaLiON Team. 2025. Meralion-audiollm: Bridging audio and language with large language models. *Preprint*, arXiv:2412.09818.

Mistral AI Team. 2025. Mistral small 3.

Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. 2025. Beyond n-grams: Rethinking evaluation metrics and strategies for multilingual abstractive summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19019–19035.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. In *Interspeech 2017*, pages 2616–2620.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. In *Interspeech 2023*, pages 4823–4827.

Hyunjong Ok and Jaeho Lee. 2025. S2cap: A benchmark and a baseline for singing style captioning. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6492–6497.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, YanGui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, Shuai Wang, and Kai Yu. 2026. A survey on speech large language models for understanding. *IEEE Journal of Selected Topics in Signal Processing*, 20(1):2–31.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Fabian Retkowski, Maike Züfle, Andreas Sudmann, Dinah Pfau, Shinji Watanabe, Jan Niehues, and Alex Waibel. 2025. Summarizing speech: A comprehensive survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27263–27294.

Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann. 2024. EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation. In *Interspeech 2024*, pages 4873–4877.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*.

Christoph Schuhmann, Robert Kaczmarczyk, Gollam Rabby, Felix Friedrich, Maurice Kraus, Kourosh Nadi, Huu Nguyen, Kristian Kersting, and Sören Auer. 2026. Emonet-voice: A fine-grained, expert-verified benchmark for speech emotion detection. *Preprint*, arXiv:2506.09827.

Hengchao Shang, Zongyao Li, Jiaxin Guo, Shaojun Li, Zhiqiang Rao, Yuanchang Luo, Daimeng Wei, and Hao Yang. 2024. An End-to-End Speech Summarization Using Large Language Model. In *Interspeech 2024*, pages 1950–1954.

StepFun. 2025. Step-audio 2 technical report. *Preprint*, arXiv:2507.16632.

Bo-Hao Su, Hui-Ying Shih, Jinchuan Tian, Jiatong Shi, Chi-Chun Lee, Carlos Busso, and Shinji Watanabe. 2026. Reasoning beyond majority vote: An explainable speechlm framework for speech emotion recognition. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 22532–22536.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. *Preprint*, arXiv:2310.13289.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, and 1 others. 2026. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Andreas Triantafyllopoulos, Anton Batliner, and Björn W. Schuller. 2026. Charting 15 years of progress in deep learning for speech emotion recognition: A replication study. *IEEE Transactions on Affective Computing*, 17(3):3185–3199.

Paige Tuttösí, Mantaj Dhillon, Luna Sang, Shane Eastwood, Poorvi Bhatia, Quang Minh Dinh, Avni Kapoor, Yewon Jin, and Angelica Lim. 2026. Bersting at the screams: A benchmark for distanced, emotional and shouted speech recognition. *Computer Speech & Language*, 95:101815.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Francesco Verdini, Pierfrancesco Melucci, Stefano Perna, Francesco Cariaggi, Marco Gaido, Sara Papi, Szymon Mazurek, Marek Kasztelnik, Luisa Bentivogli, Sebastien Bratières, Paolo Merialdo, and Simone Scardapane. 2025. How to Connect Speech Foundation Models and Large Language Models? What Matters and What Does Not. In *Interspeech 2025*, pages 1813–1817.

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. 2025a. Capspeech: Enabling downstream applications in style-captioned text-to-speech. *Preprint*, arXiv:2506.02863.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Qiongqiong Wang, Hardik Bhupendra Sailor, Tianchi Liu, Wenyu Zhang, Muhammad Huzaifah, Nattadaporn Lertcheva, Shuo Sun, Nancy F. Chen, Jinyang Wu, and AiTi Aw. 2025b. 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*, pages 14133–14148, Suzhou, China. Association for Computational Linguistics.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024a. Secap: Speech emotion captioning with large language model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19323–19331.

Yaoxun Xu, Yixuan Zhou, Yunrui Cai, Jingran Xie, Runchuan Ye, and Zhiyong Wu. 2024b. Multimodal emotion captioning using large language model with prompt engineering. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pages 104–109.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive:(http://web. ku. edu/˜ idea/readings/rainbow. htm)*.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, arXiv:2505.09388.

Shu-wen Yang, Ming Tu, Ting-Wei Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. 2026. Paras2s: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. In *International Conference on Learning Representations*, volume 2026, pages 70704–70729.

Linhao Zhang, Jian Zhang, Bokai Lei, Chuhan Wu, Aiwei Liu, Wei Jia, and Xiao Zhou. 2025a. Wildspeech-bench: Benchmarking end-to-end speechllms in the wild. *Preprint*, arXiv:2506.21875.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zihan Zhang, Shreeram Suresh Chandra, Zongyang Du, Xiutian Zhao, Aurosweta Mahapatra, Hao Zhang, Philipp Koehn, and Berrak Sisman. 2026. Trace-evc: Text-guided relative affective control for zero-shot emotional voice conversion. *Preprint*, arXiv:2607.03666.

Zixing Zhang, Yimeng Wu, Zhongren Dong, Wulong Xiang, Shengfan Shen, and Björn W. Schuller. 2025b. Sse: A speaking style extractor based on fine-grained contrastive learning between speech and descriptive text. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Zixing Zhang, Weixiang Xu, Zhongren Dong, Kanglin Wang, Yimeng Wu, Jing Peng, Runming Wang, and Dong-Yan Huang. 2024. Paralbench: A large-scale benchmark for computational paralinguistics over acoustic foundation models. *IEEE Transactions on Affective Computing*.

Xiutian Zhao, Ismail Rasim Ulgen, Philipp Koehn, Björn Schuller, and Berrak Sisman. 2026. Neuron-level emotion control in speech-generative large audio-language models. *Preprint*, arXiv:2603.17231.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in neural information processing systems*, volume 36, pages 46595–46623.

Xinfa Zhu, Wenjie Tian, Xinsheng Wang, Lei He, Yujia Xiao, Xi Wang, Xu Tan, Sheng Zhao, and Lei Xie. 2024. Unistyle: Unified style modeling for speaking style captioning and stylistic speech synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 7513–7522, New York, NY, USA. Association for Computing Machinery.

A Existing Datasets and Benchmarks

Name	Size (hrs)
ParlerTTS (Lyth and King, 2024)	45k
PromptSpeech (Guo et al., 2023)	37 (real)
Expresso (Nguyen et al., 2023)	47
EARS (Richter et al., 2024)	60
TextrolSpeech (Ji et al., 2024)	300
MEAD-TTS (Guan et al., 2024)	36
SpeechCraft (Jin et al., 2024)	2.4k
LibriTTS-P (Kawamura et al., 2024)	600
Dreamvoice (Hai et al., 2024)	231
Capspeech (Wang et al., 2025a)	33.6k
ParaSpeechCaps (Diwan et al., 2025)	2.7k

Table 5: Summary of existing datasets relevant to SSC.

B Annotation reliability of PSC style metadata

The quality of PSC annotations is supported by multiple stages of annotation verification and filtering throughout the dataset construction pipeline. For PSC-Base, annotations were collected from qualified Amazon Mechanical Turk annotators who completed a dedicated qualification task designed around identifying speaking-style attributes, and annotations were aggregated across five annotators per speaker with agreement-based filtering applied before retaining tags. For PSC-Scaled, automatic annotation combines complementary components including perceptual speaker similarity, expressivity filtering, semantic matching, and acoustic matching using an audio LLM-as-a-judge, rather than relying on a single classifier or heuristic. Importantly, the paper further verifies scaled data quality through human evaluation, showing that PSC-Scaled achieves tag recall comparable to PSC-Base for both intrinsic and situational tags, supporting that the automatically scaled annotations remain aligned with human perceptual judgments.

Benchmark	What Task It Benchmarks	Evaluation Metrics
CP-Bench (Wang et al., 2025b)	Contextual Paralinguistic Reasoning: Evaluates the model's ability to reason over both verbal content and non-verbal cues like emotion and prosody to answer questions	Answer accuracy, human preference ratings
SD-Eval (Ao et al., 2024)	Spoken Dialogue Understanding: assesses understanding of dialogue that relies on paralinguistic cues (e.g., sarcasm, tone) for correct interpretations	Task accuracy, BLEU (for generation), MOS-style human ratings.
ParaS2SBench(Yang et al., 2026)	Paralinguistic-Aware Speech-to-Speech: Evaluates if generated speech responses maintain appropriate paralinguistic alignment (style/emotion) with the input	Content similarity, style similarity, MOS (Mean Opinion Score)
ParaLBench (Zhang et al., 2024)	General Computational Paralinguistics: A comprehensive test suite across multiple existing paralinguistic datasets to test generalizability	Accuracy, UAR (Unweighted Average Recall), F1 score, correlation metrics
BERSt (Tuttösí et al., 2026)	Expressive Emotion Recognition: Specifically benchmarks emotion recognition in challenging conditions like shouted or highly expressive speech	SER (Speech Emotion Recognition) accuracy, UAR, WER (Word Error Rate - as a stress test)
EmoNet-Voice-Bench (Schuhmann et al., 2026)	Fine-grained Emotion Understanding: Tests understanding of nuanced emotions beyond basic categories (e.g., distinguishing "cold anger" from "hot anger")	Classification accuracy, macro-F1, confusion matrix analysis
StepEval-Audio-Paralinguistic (StepFun, 2025)	Multidimensional paralinguistic comprehension: Evaluates 11 distinct dimensions including gender, age, pitch, rhythm, and speaking rate.	Dimension-wise accuracy, human ratings
S2Cap (Ok and Lee, 2025)	Singing Style Description: Focuses on describing singing voices using natural language, capturing attributes like mood, tempo, and timbre	Classification accuracy on style factors, Captioning metrics: BLEU, METEOR, CIDEr
SpeechCaps (Huang et al., 2025)	Speaking-Style Captioning in multi-talker scenarios: Benchmarks the ability to generate captions that describe speaker-specific styles and prosody in multi-talker scenarios.	Instruction-following rate, overall accuracy, and conditional accuracy, with correctness and relevance judged via LLM-based semantic alignment (GPT-4o)
WildSpeech-Bench (Zhang et al., 2025a)	Natural / In-the-Wild Prosody: Benchmarks how models handle real-world paralinguistic variability, including stuttering, hesitations, interruptions, and extreme prosodic conditions (e.g., whispering vs. shouting).	UGPT-4o-based 1-10 scoring, human-alignment metrics (Pearson/R^2/MSE), and UTMOS.
Multi-Bench (Deng et al., 2025)	Multi-turn Emotional Dialogue: Tests whether models can maintain emotional consistency and empathy over a long conversation, unlike single-turn benchmarks.	Consistency score, empathy rating (Likert scale by human or LLM judge)
Salmon (Maimon et al., 2025)	Acoustic consistency and semantic-acoustic alignment across attributes such as sentiment, speaker identity, background noise, and room acoustics using likelihood-based comparisons.	Uses pairwise likelihood preference scores (accuracy) for acoustic consistency and semantic-acoustic alignment tasks, with sWUGGY for spoken-content modeling and human agreement as an upper bound.

Table 6: Benchmarks for Speech Paralinguistics Understanding

C StyleInstructCaps-DB Construction details


```
{
"source": "ears",
"relative_audio_path": "p047/emo_amazement_freeform.wav",
"text_description": [
" A female speaker's authoritative, nasal voice is crisp and high-pitched, conveyed in an American
accent with a measured speed. Her speech is delivered in a slightly clean environment, yet she
expresses awe with a loud and slightly elevated volume.",
" An American woman speaks with a high-pitched voice, expressing awe, in a slightly clean
environment. Her speech is measured and deliberate."
],
"transcription": " Oh, really? You were so good at that. I love to just sit here and listen to you
play. It's so relaxing. I just can't believe how talented you are.",
"intrinsic_tags": [
"authoritative",
"crisp",
"loud",
"american",
"nasal"
],
"situational_tags": [
"awed"
],
"basic_tags": [
"female",
"high-pitched",
"measured speed",
"slightly clean environment"
],
"all_tags": [
"american",
"authoritative",
"awed",
"crisp",
"female",
"high-pitched",
"loud",
"measured speed",
"nasal",
"slightly clean environment"
],
"speakerid": "p047",
"name": "p047",
"duration": 9.8,
"gender": "female",
"accent": "american",
"pitch": "high-pitched",
"speaking_rate": "measured speed",
"noise": "slightly clean environment",
"utterance_pitch_mean": 220.68133544921875,
"snr": 56.96498489379883,
"phonemes": " oʊ, ɹɪli? ju wɝ soʊ ɡʊd æt ðæt. aɪ lʌv tu dʒʌst sɪt hiɹ ʌnd lɪsʌn tu ju pleɪ. ɪt'ɛs soʊ
ɹɪlæksɪŋ. aɪ dʒʌst kæn'ti bɪliv haʊ tælʌntɪd ju ɑɹ.",
"audio_path": "datasets/ears/p047/emo_amazement_freeform.wav",
}
```


Figure 4: Example metadata object from EARS used in our captioning/evaluation pipeline.

```
[ROLE & GOAL]
You are an AI expert in speech prosody. Your primary task is to generate six highly descriptive, fine-grained
speech captions from a given audio metadata object.

[CONTEXT: INPUT METADATA]
Analyze the following JSON object, which contains detailed metadata for a speech utterance.

JSON
{metadata}

[TASK & INSTRUCTIONS]
Based on the metadata, generate six unique and descriptive captions. For each caption, combine multiple vocal
characteristics using descriptive verbs and adverbs to precisely capture the essence of the speaking-style.
Use these general guidelines for each caption:

1. Think in Verbs and Adverbs: Instead of "sad," think "a voice that wavers and falls off slowly."
2. Combine at Least Three Elements: A good caption should ideally define a pitch contour, a pace, and a vocal
quality.
3. Specify the Context: Adding a phrase like "as if explaining a complex topic" or "like a conspiratorial
whisper" can give the model a powerful contextual anchor.
4. Use the Tags: Leverage the intrinsic_tags, situational_tags, basic_tags, and all_tags to inform your
captions.
5. Always associate the nationality with the accent
6. Use the provided transcription to infer emotional and contextual elements but do not directly quote it in
the captions.

Each caption should focus on a different aspect of the speech, as outlined below.

1. Speaker-Idiosyncratic Style:
Goal: Capture the speaker's core vocal identity.
Instruction: Generate a caption describing the speaker's inherent qualities. Combine the pitch , accent, gender
, and timbre-related intrinsic_tags.

2. Situational/Contextual Style:
Goal: Describe the circumstances of the speech.
Instruction: Generate a caption that sets the scene. Combine the speaking_rate (measured speed), the noise
(noisy environment), and contextual intrinsic_tags (authoritative). Frame it as a person speaking with purpose
in a specific setting.

3. Expressive/Emotional Style:
Goal: Convey the inferred emotion or attitude.
Instruction: Generate a caption describing the emotional tone. Infer the attitude from the transcription,
situational_tags and all_tags. Describe the prosody (pitch contour, rhythm) that would create this feeling.

4. Linguistic/Pragmatic Style:
Goal: Detail the speech's structural and emphasis patterns.
Instruction: Generate a caption focused on the delivery of the words themselves. Analyze the transcription to
identify linguistic features. Describe the pausing and stress patterns.

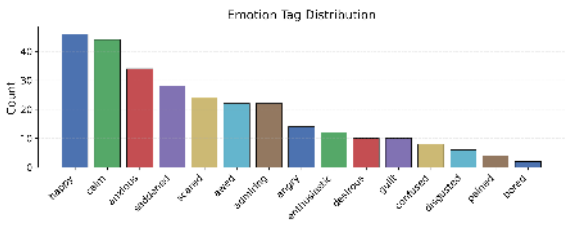
5. Perceptual/Listener-Centric Style:
Goal: Describe the impression the voice leaves on a listener.
Instruction: Generate a caption from the listener's point of view. Use impressionistic intrinsic_tags to
describe the overall effect of the voice. How does it feel to hear this person speak?

6. Overall style summary:
Goal: Create a rich, narrative-style caption.
Instruction: Combine all elements in instrinsic-tags, situational_tags, basic_tags, all_tags, gender,
accent, pitch, speaking_rate, noise into a creative, narrative-style caption that captures the essence of
the speaking-style.

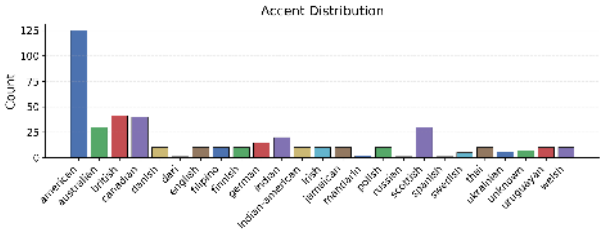
[OUTPUT FORMAT & CONSTRAINTS]
Return your response as a single, valid JSON object. Use the descriptive keys provided in the example below.
The entire output must be only the JSON object, with no additional text, explanations, or markdown formatting.

JSON
{
"speaker_idiosyncratic_style": "<Generated caption for style 1>",
"situational_contextual_style": "<Generated caption for style 2>",
"expressive_emotional_style": "<Generated caption for style 3>",
"linguistic_pragmatic_style": "<Generated caption for style 4>",
"perceptual_listener_centric_style": "<Generated caption for style 5>",
"overall_style_summary": "<Generated caption for style 6>"
}
```

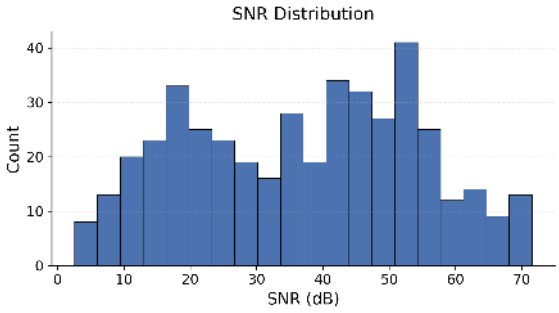
Figure 5: Prompt template for generating six style-specific captions from an input metadata object.



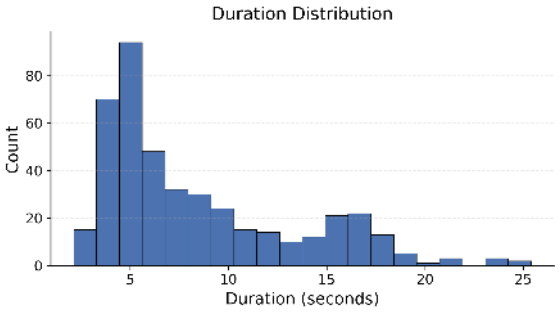
(a) Emotion



(b) Accent

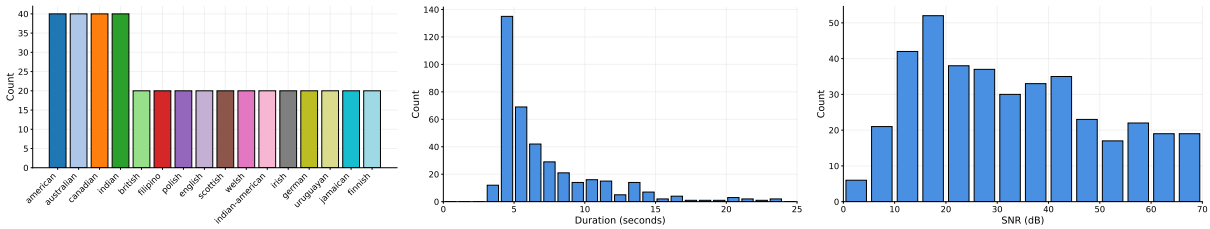


(c) SNR



(d) Duration

Figure 6: Distributions of emotion (a), accent (b) signal-to-noise (SNR) ratio (c), and speech duration (d) in *StyleInstructCaps-MetaSet*.



(a) Accent (b) Duration (c) SNR

Figure 7: Distributions of (a) accent (b) duration and (c) signal-to-noise (SNR) ratio in StyleInstructCaps In-domain balanced set (ID-Balanced).

Instruction type	Style instruction	Generated caption
Speaker idiosyncratic style	Describe the speaker's inherent vocal identity by combining their pitch, accent, gender, and timbre-related characteristics.	A high-pitched, crisp, and nasal voice with an American accent, delivered by a female speaker, instantly conveys a sense of authority and confidence.
Situational contextual style	Set the scene by combining the speaking rate, noise level, and contextual intrinsic tags to describe a person speaking with intention in a specific setting.	Speaking with measured speed and authority, a female voice with an American accent cuts through a slightly clean environment, commanding attention as if addressing a gathering with purpose and conviction.
Expressive emotional style	Infer the emotional tone from the transcription and situational tags, and describe the prosody that creates this feeling, capturing the speaker's attitude and sentiment.	A voice filled with awe and admiration, conveyed through a rising pitch contour and deliberate pacing, expresses genuine astonishment, as if the speaker is witnessing something remarkable for the first time.
Linguistic pragmatic style	Analyze the transcription to identify linguistic features, and describe the pausing and stress patterns that reveal the speaker's emphasis and intention.	The speech unfolds with a steady, measured rhythm, punctuated by deliberate pauses that emphasize key phrases, as the speaker carefully chooses words to convey their thoughts, creating a sense of sincerity and authenticity.
Perceptual listener-centric style	Describe the overall effect of the voice on the listener, using impressionistic intrinsic tags to capture the impression the voice leaves on those who hear it.	Listening to this voice is like being drawn into a conversation with a charismatic and confident individual, as the speaker's authoritative tone and crisp delivery create a sense of trust and engagement.
Holistic creative synthesis	Combine all elements to create a rich, narrative-style caption that captures the essence of the speaking-style, weaving together the speaker's inherent qualities, situational context, emotional tone, and linguistic features into a vivid and immersive portrait.	Imagine a female speaker with a high-pitched, crisp, and nasal voice, speaking with measured speed and authority, her American accent shining through as she expresses awe and admiration in a slightly clean environment, conveying confidence and charisma with every word.

Table 7: Style-specific instructions and example generated captions.

Table 8: Summary of dataset splits.

Data split	Number of utterances	Number of hours
PSC/SIC train set	1031840	2681.54
PSC/SIC validation set	11967	26.29
PSC/SIC evaluation set	14756	33.04
StyleInstructCaps-MetaSet	434	1.02
In-domain balanced set	400	0.82
Unseen Data(UD) - VCTK	320	0.31
VCTK (unseen prompts)	32	0.03
Gigaspeech (with SpeechCraft captions)	600	1.00

D LLM- judge calibration : Adversarial gaming check

The adversarial gaming check is designed as a controlled stress-test for the LLM-as-a-judge evaluation pipeline used in StyleInstructCaps. Starting from the 434 utterances in StyleInstructCaps-MetaSet, we construct adversarial captions that are intentionally fluent, instruction-aligned, or stylistically convincing while violating grounding constraints in different ways. Each attack category targets a specific potential failure mode of the judge, such as rewarding verbosity, surface-level instruction alignment, or stylistic fluency despite factual inconsistencies with the underlying metadata. We use the gemma2-27B (Team et al., 2024) to generate the adversarial captions. The adversarial gaming check is performed for a single-run of the Qwen3-32B judge.

The resulting evaluation setup enables us to measure whether the judge meaningfully penalizes adversarial behavior rather than assigning inflated scores to superficially plausible captions. Table 9 summarizes the five adversarial attack categories, the evaluation dimensions they target, and the corresponding failure modes being tested.

For each attack type, we compare judge scores on oracle vs. adversarial captions using a two-tailed (paired t-test), and report effect size as Cohen's d. The results of the adversarial calibration experiment, as visualised in Figure 8 demonstrates that the Qwen-3-32B LLM-as-a-judge is highly sensitive to multiple forms of adversarial caption manipulation and does not merely reward fluent or stylistically convincing text. Across all three evaluation axes, oracle captions consistently achieve near-ceiling scores, whereas adversarial captions incur statistically significant degradations.

For the metadata groundedness (Meta-Score) evaluation, the judge consistently penalizes all five attack types. Oracle captions obtain very high scores (mean $\sim$ 9.5), while adversarial captions exhibit substantial drops across all conditions. The most pronounced degradations are observed for Type 4 (Hedged Hallucination) and Type 5 (Cross-Speaker Swap), with score gaps of 7.46 and 6.41 respectively and very large effect sizes ($d = 3.82$ and $d = 7.26$, both $p < 0.001$). Even the weakest attack, Type 1 (Attribute Inversion), yields a statistically significant degradation ($d = 1.39$, $p < 0.001$). These findings indicate that the judge is strongly grounded in metadata consistency and remains robust to both explicit and implicitly hedged factual violations.

The instruction-following (IFA) evaluation further demonstrates that the judge performs an independent assessment rather than collapsing instruction adherence with metadata correctness. Type 2 (Instruction Parroting) induces the largest degradation in IFA performance, reducing scores from approximately 9.0 to 2.59 with a large effect size ($d = 4.23$, $p < 0.001$). Type 4 (Hedged Hallucination) also significantly degrades IFA performance ($d = 3.73$, $p < 0.001$). In contrast, Type 1 (Attribute Inversion) has only a minor impact on IFA ($d = 0.34$, $p < 0.05$), as the captions preserve structural compliance with the instruction despite factual corruption. Type 3 (Verbosity Masking) yields a moderate reduction ($d = 1.58$, $p < 0.001$), indicating that stylistically rich but semantically ungrounded outputs are not rewarded as valid instruction-following instances. Overall, these results show that the judge disentangles instruction adherence from factual grounding.

The hallucination severity analysis reveals a more nuanced pattern. Type 1 (Attribute Inversion) and Type 5 (Cross-Speaker Swap) consistently produce the highest hallucination rates, confirming that explicit factual contradictions are reliably detected. Type 2 (Instruction Parroting) is also strongly penalized despite containing limited explicit factual content, suggesting that the absence of grounded information itself contributes to hallucination judgments. Interestingly, Type 4 (Hedged Hallucination) is not consistently assigned high hallucination severity, despite being strongly penalized under Meta-Score and IFA evaluations. This indicates that epistemically hedged formulations may partially obscure hallucination detection under the dedicated hallucination rubric, even though they remain identifiable as low-quality or weakly grounded under other evaluation axes.

Additionally, we observe that oracle captions remain near-saturated across metrics, supporting evaluation stability; SALMONN outputs score below oracle references but well above adversarial captions, indicating the judge does not penalize all non-oracle outputs.

Overall, these results provide strong evidence that the LLM-as-a-judge framework is robust to a diverse set of adversarial manipulations. The judge consistently distinguishes grounded from ungrounded captions, remains sensitive to factual inconsistencies even under hedging, and evalu-

Attack Type	Primary Target Metric	Failure Mode Tested	Count
Type 1: Attribute Inversion	Meta-Score, Hallucination	The caption preserves fluency and overall structure while directly contradicting core metadata attributes such as gender, accent, or pitch. This tests whether the judge anchors on stylistic fluency rather than factual correctness with respect to the provided metadata.	87
Type 2: Instruction Parroting	Instruction-Following Accuracy (IFA)	The caption directly restates the instruction as if it were a caption, without providing any grounded acoustic or stylistic content. This tests whether the judge rewards instruction-aligned surface form rather than true instruction compliance.	87
Type 3: Verbosity Masking	Meta-Score, Hallucination, IFA	The caption uses long, fluent, impressionistic prose filled with prosodic language while deliberately avoiding verifiable factual claims. This tests whether the judge exhibits verbosity bias by inflating scores for stylistically rich but ungrounded captions.	87
Type 4: Hedged Hallucination	Hallucination Severity	The caption introduces incorrect factual claims through epistemically hedged language such as "may suggest" or "could be interpreted as." This tests whether the judge is overly lenient toward softened or non-committal hallucinations.	87
Type 5: Cross-Speaker Swap	Meta-Score	The caption uses a real oracle caption from a different speaker with mismatched demographic and emotional attributes. This tests whether the judge grounds evaluation in the provided metadata or simply rewards any fluent, high-quality caption.	86

Table 9: Summary of the five adversarial attack categories used in the LLM judge calibration experiment. Each attack probes a distinct potential failure mode of the LLM-as-a-judge evaluation pipeline.



Figure 8: Calibration analysis of speaking-style captioning models across metadata groundedness, instruction-following and hallucination severity metrics.

ates instruction-following, hallucination, and metadata consistency as partially decoupled dimensions rather than conflating them into a single fluency-based score.

E Human-LLM judge correlation study

Table 1 ranks speaking-style captioning systems with scores from a panel of LLM judges. An LLM judge can reward fluent or detailed text that listeners do not hear in the speech. We therefore test the judge against human listeners. The study answers four questions, each addressed by one results paragraph below. **(Q1)** Do listeners agree with each other well enough to serve as a reference? **(Q2)** Does the judge prefer the same caption as listeners, pair by pair? **(Q3)** Is the judge more often right when its score gap between two captions is larger? **(Q4)** Does the system ranking survive when it is fit to human votes alone?

E.1 Study design

Judge score under test. For each caption c, the panel assigns a metadata-groundedness score $s(c)$ on a 1–10 scale (Section 4.1). The *ensemble* score is the mean over three judges (Qwen3-32B, Gemma-4-31B-it, GPT-OSS-20B) and three runs per judge, nine runs in total (Appendix F). This is the score reported in Table 1. The judge reads the utterance metadata and the caption and never hears the audio. Listeners hear the audio and never see the metadata. The study therefore tests whether a score computed from metadata agrees with what listeners perceive in the speech.

Caption pairs and score gap. Each of the 100 study items consists of one utterance from *StyleInstructCaps-MetaSet* and the captions of two different systems for that utterance. The 14 systems include the oracle captions of *StyleInstructCaps-DB*, which enter the study as one system alongside the 13 captioning models (the 12 open-weight models and GPT-Audio); the PSC oracle is not included. For an item with captions c_a and c_b, the signed score gap is $\Delta = s(c_a) - s(c_b)$. The judge prefers c_a when $\Delta > 0$ and c_b when $\Delta < 0$. The absolute gap $|\Delta|$ measures how strongly the judge separates the two captions. A near-zero gap means the judge rates both captions almost equally. Items were sampled to cover all 14 systems defined above and the full range of $|\Delta|$. Items with small $|\Delta|$ were over-sampled. These are the decisions on which a ranking error is most likely, so the study spends most of its trial budget there instead of on easy pairs. For analysis, we partition the observed range of $|\Delta|$ into five equal-width buckets, from B1 (smallest gaps, *near-tie*) to B5 (largest gaps).

Task. On each trial, a listener heard the raw audio clip, read the two captions, and chose the one that better described the speaking style. The order of the two captions was randomized. We use a binary forced choice for two reasons. It matches how the benchmark is used, which is to decide which of two systems is better. It also avoids absolute 1–10 ratings, whose scale drifts from listener to listener. All listeners rated the same 100 items, so every agreement statistic below is computed on a fully crossed design.

Listeners and quality control. Fifteen English-speaking participants took part. Eight attention-check trials with a known correct answer were mixed into the session. A listener had to answer at least 6 of the 8 checks correctly (75%) to be retained. Two listeners failed, with 3/8 (37.5%) and 5/8 (62.5%) correct, and are excluded. All results use the remaining 13 listeners and their 1,300 votes. For each item, the *human majority* is the caption chosen by at least 7 of the 13 listeners. With an odd number of listeners, the majority is always defined.

E.2 Metrics

Agreement among listeners (Q1). Krippendorff's α and Fleiss' κ measure how often listeners pick the same caption, corrected for the agreement expected by chance. A value of 0 means chance-level agreement and 1 means perfect agreement. Mean pairwise agreement is the uncorrected fraction of items on which two listeners pick the same caption, averaged over all 78 listener pairs. Its chance level is 50%. We also count how evenly each item's 13 votes divide. An item is *unanimous* when all 13 listeners pick the same caption. Non-unanimous items are *lopsided* when one caption wins by a wide margin and *split* when the vote is close to even.

Agreement between judge and listeners (Q2). *Agreement* is the fraction of items on which the judge prefers the same caption as the human majority. Items on which the judge assigns both captions the same score ($\Delta = 0$) have no judge preference and are excluded. The number of remaining items n is reported per judge. Cohen's κ corrects agreement for chance, with the same 0-to-1 reading as above. The p-value is a two-sided exact binomial test of agreement against the 50% expected from

Table 10: Inter-listener reliability and quality control for the human pairwise study (13 retained listeners, 100 items). Chance level for pairwise agreement is 50%.

Measure	Value
Krippendorff's α / Fleiss' κ	0.347 / 0.346
Mean pairwise agreement	67.3%
Split / lopsided / unanimous items	45 / 44 / 11
Attention-check pass rate (all 15 listeners)	83.3%

a judge that guesses. All intervals are 95% confidence intervals.

Calibration against the score gap (Q3). If $|\Delta|$ reflects how confident the judge is, agreement should rise with $|\Delta|$. We test this in two ways. First, we report agreement per $|\Delta|$ bucket. Second, we fit a logistic regression to the 1,300 individual votes (top), and we fit a Bradley–Terry model for Q4 (bottom):

$$P(c_a \succ c_b) = \sigma(\beta_0 + \beta_1 \Delta),$$
$$P(i \succ j) = \frac{\pi_i}{\pi_i + \pi_j}.$$

Here σ is the logistic function and c_a is the first-presented caption. A positive slope β_1 means that listeners prefer the higher-scored caption more often as the gap grows. The magnitude of the score then carries information, not only its sign. The intercept $\sigma(\beta_0)$ is the probability of choosing the first-presented caption when the judge sees a tie. A value near 50% indicates no position bias.

System-level ranking (Q4). The Bradley–Terry model assigns each system i a strength π_i and gives the probability that system i's caption is preferred over system j's. We fit it by maximum likelihood to all 1,300 individual human votes, without any judge scores, and rank systems by π_i. We compare this ranking over the 14 systems with the system ranking induced by each judge's metadata-groundedness scores. Spearman's ρ correlates the two rank orders. Kendall's τ is the fraction of system pairs ordered the same way minus the fraction ordered differently, so it penalizes every local swap.

E.3 Results

Q1: Listeners agree above chance, and the majority vote is the reference. Table 10 reports agreement among the 13 retained listeners. Krippendorff's α is 0.347 and Fleiss' κ is 0.346. Two listeners pick the same caption on 67.3% of items, against a 50% chance level. Individual listeners therefore often disagree. This is expected for two reasons. Judging which caption better describes a speaking style is subjective, and the sampling deliberately over-represents near-tie items. Accordingly, 45 of the 100 items received a split vote. For this reason, we compare the judge with the majority of 13 listeners and not with any single listener. The majority vote averages out individual variation.

Table 11: Agreement between the human majority vote and the preference of each LLM judge and of the reported ensemble. Items on which a judge assigns equal scores are excluded, leaving n items. Brackets give 95% confidence intervals. p tests agreement against a 50% chance level.

Judge	n	Agreement (%)	Cohen's κ	p
Ensemble	99	85.9 [77.8, 91.9]	0.717	1×10^{-13}
Qwen3-32B	65	83.1 [73.8, 92.3]	0.661	6×10^{-8}
GPT-OSS-20B	96	82.3 [74.0, 89.6]	0.646	1×10^{-10}
Gemma-4-31B-it	92	84.8 [76.1, 91.3]	0.696	6×10^{-12}

Table 12: Agreement between the human majority and the ensemble judge as a function of the absolute ensemble score gap $|\Delta|$, partitioned into five equal-width buckets. p is a two-sided exact binomial test against 50%.

Bucket	n	Agreement	p
B1 (near-tie)	15	73.3%	0.12
B2	20	80.0%	0.012
B3	24	87.5%	3×10^{-4}
B4	25	92.0%	2×10^{-5}
B5 (large gap)	15	93.3%	1×10^{-3}

Q2: The judge prefers the same caption as listeners on most items. Table 11 reports the results. The ensemble agrees with the human majority on 85.9% of the 99 items it decides, with Cohen's $\kappa = 0.717$ ($p = 1 \times 10^{-13}$). Individual judges reach 82.3% to 84.8%. The ensemble has the highest point estimate, but its confidence interval overlaps with those of the individual judges, so we do not claim a significant gain from ensembling. The ensemble averages nine runs and assigns tied scores less often, so it decides more items than any single judge. Excluding the 15 near-tie items (B1) leaves 84 items, on which agreement rises to 88.1% (95% CI [81.0, 95.2]).

Q3: Agreement rises with the judge's score gap. Table 12 reports agreement per bucket. Agreement rises monotonically from 73.3% in the near-tie bucket to 93.3% in the largest-gap bucket. Buckets B2 to B5 are each significantly above chance ($p \leq 0.012$). The near-tie bucket is not ($p = 0.12$, 11 of 15 items). On near-tie items the judge rates

Table 13: Rank correlation over the 14 systems between the ranking of a Bradley–Terry model fit to human votes alone and the ranking produced by each LLM judge.

Judge	Spearman ρ	Kendall τ
Ensemble	0.895	0.780
Qwen3-32B	0.873	0.707
GPT-OSS-20B	0.855	0.692
Gemma-4-31B-it	0.881	0.714

both captions almost equally, and listeners split in the same way. The size of a score difference in Table 1 therefore predicts how reliably listeners share the judge's preference.

Q4: Human votes alone recover the judge's system ranking. Table 13 compares the two rankings. The ensemble reaches Spearman $\rho = 0.895$ and Kendall $\tau = 0.780$. Individual judges reach ρ between 0.855 and 0.881. The five systems ranked lowest by listeners are the same five ranked lowest by the judge: Audio-Flamingo-3, Qwen2-Audio-7B-Instruct, MERaLiON-AudioLLM, Voxtral-Small-24B and SALMONN. At the top of the ranking, the human-fit model places MOSS-Audio above the oracle captions. The 95% confidence intervals of the two Bradley–Terry strengths overlap heavily, so listeners do not separate these two systems.

Summary. Listeners agree with each other above chance, and their majority vote provides a reference for the judge. The judge panel prefers the same caption as the majority on 85.9% of items. Its agreement grows with its own score gap, so score magnitudes carry information that listeners confirm. A ranking fit to human votes alone reproduces the judge's system ranking at $\rho = 0.895$. These results support the use of the LLM judge panel for the caption-quality scores reported in Table 1.

Table 14: LLM judges used in the evaluation panel.

Judge	Family / size	Role
Qwen3-32B (thinking)	Qwen, 32B dense	Panel judge; instruction-following
Gemma-4-31B-it (thinking)	Gemma, 31B dense	Panel judge
GPT-OSS-20B (reasoning)	OpenAI OSS, 20B MoE	Panel judge

Table 15: Number of runs per judge and metric.

Evaluation set	Meta. groundedness	Hallucination	Instr.-following
MetaSet	3× each judge	3× each judge	1× Qwen3-32B
OOD (GigaSpeech, VCTK)	2× each judge	2× each judge	1× Qwen3-32B

F Panel of LLM judges: cross-judge agreement

To test whether the benchmark conclusions depend on the choice of evaluator, we evaluate the open-ended metrics with three independent open-weight LLM judges and execute each judge multiple times. This design quantifies cross-judge agreement and the stability of system rankings across judges.

Judge panel. The panel consists of three open-weight judges spanning different model families and scales: Qwen3-32B, Gemma-4-31B-it, and GPT-OSS-20B (Table 14). All judges receive the same prompts (Appendix H). Table 15 lists the number of runs per judge and metric. Instruction-following is scored with a single Qwen3-32B run because it requires 2,604 judgments per audio-language model, judge, and run.

Cross-judge agreement. Because the benchmark is used to rank systems, the central question is whether the ranking is an artifact of a specific judge. We compute Spearman's rank correlation between the per-judge mean metadata-groundedness scores of the 14 benchmarked systems on *StyleInstructCaps-MetaSet* (Table 16). All three judge pairs agree strongly ($\rho \geq 0.864$), and Gemma-4-31B-it and GPT-OSS-20B agree almost perfectly ($\rho = 0.978$).

Reproducibility. We release the judge model identifiers, prompts, and decoding settings for all three judges together with the benchmark.

Table 16: Cross-judge agreement: Spearman rank correlation between judges over the 14 benchmarked systems (*MetaSet* metadata groundedness, per-judge means).

Judge pair	Spearman ρ ($n = 14$)
GPT-OSS-20B ↔ Gemma-4-31B-it	0.978
Qwen3-32B ↔ GPT-OSS-20B	0.921
Qwen3-32B ↔ Gemma-4-31B-it	0.864

Table 17: Omnibus Friedman test and number of significantly different system pairs ($p < 0.05$) out of all $\binom{n}{2}$ pairs, per metric and evaluation set. n is the number of systems compared. The *MetaSet* groundedness and hallucination tests use all 15 systems of Table 1 (two oracle caption sets, 12 open-weight models, and GPT-Audio). Instruction-following excludes the PSC oracle and GPT-Audio, which have no instruction-following scores. The OOD groundedness and hallucination tests use the 13 systems of Table 3. The large majority (86–96%) of pairs are statistically distinguishable; the remaining pairs are statistically close.

Eval set (table)	Metric	n	Friedman p	Sig. pairs	%
MetaSet (Tab. 1)	Groundedness ↑	15	$< 10^{-4}$	101/105	96.2
	Hallucination ↓	15	$< 10^{-4}$	99/105	94.3
	Instr.-following ↑	13	$< 10^{-4}$	73/78	93.6
OOD (Tab. 3)	Groundedness ↑	13	$< 10^{-4}$	68/78	87.2
	Hallucination ↓	13	$< 10^{-4}$	67/78	85.9
OOD (Tab. 4)	Instr.-following ↑	14	$< 10^{-4}$	79/91	86.8

G Pairwise statistical significance of system rankings

Method. Our results tables rank systems against each other. We therefore test whether the gap between two systems reflects a real difference or only noise from the judges or the utterances. We first compute a per-utterance consensus score for each system, which gives one score per system for every utterance and lets us pair scores across systems. For metadata groundedness and hallucination, this score is the mean over the system's panel runs (9 for *MetaSet* and 6 for the out-of-domain set; Appendix F). For instruction-following, it is the score from the single Qwen3-32B run. We then compare every pair of systems with a two-tailed paired test over utterances. We use both a paired t-test and a Wilcoxon signed-rank test, and the two agree in every case. Groundedness and hallucination scores are paired by utterance index, and instruction-following scores are paired by (utterance, instruction type). We report significance at $p < 0.05$ along with effect sizes, namely the mean paired difference Δ and Cohen's d. Before looking at individual pairs, we run an omnibus Friedman test to check whether the systems differ at all on a given metric.

Systems differ, and the leaderboards are statistically well separated. The Friedman test is highly significant ($p < 10^{-4}$) for every metric and every evaluation set. Table 17 shows how many of the $\binom{n}{2}$ system pairs differ at $p < 0.05$.

Our central claims are statistically significant. Table 18 shows that the paired tests confirm the key findings of the paper on *MetaSet* metadata groundedness (Table 1). Our main claim is that SSC-specific, instruction-diverse training improves groundedness, and this gain is statistically significant. The oracle upper bound scores significantly higher than every real model, which validates the metric. The closed-source GPT-Audio scores significantly lower than the best open model on groundedness and significantly worse on hallucination on both evaluation sets (all $p < 10^{-4}$). This is consistent with our finding that following instructions does not imply grounded understanding.

Near-identical systems are not separated. The tests do not distinguish systems that are nearly identical by design. On the out-of-domain set, for example, the four SALMONN variants (means 6.25–6.30) do not differ significantly from one another (SSC-MT vs. baseline SALMONN: $\Delta = +0.02$, $p = 0.69$). The non-significant pairs in Table 17 are near-ties of this kind.

Table 18: Headline comparisons on *MetaSet* metadata groundedness (↑; $N = 434$ utterances). Δ is the mean per-utterance score gap. A Cohen's d above 0.8 is conventionally a large effect, so these gaps are both significant and sizeable.

Comparison	Δ	d	Wilcoxon p
SSC-tuned SALMONN (SIC-DB, MT) > baseline SALMONN	$+2.67$	1.51	$< 10^{-4}$
SSC-tuned SALMONN (SIC-DB, ST) > baseline SALMONN	$+2.62$	1.53	$< 10^{-4}$
Oracle (SIC-DB) > best real model	$+2.51$	1.70	$< 10^{-4}$
Best open model > GPT-Audio (closed SOTA)	$+1.65$	0.87	$< 10^{-4}$

H Prompts used

```
You are an expert linguist and speech analyst tasked with comparing the speaking-styles between two sets of audio metadata. Your goal is to determine the similarity between a ground truth recording and a generated caption, focusing particularly on stylistic aspects of speech, intention, and prosody.

Please carefully review the following metadata:

Ground Truth Metadata (DATA1):
<ground_truth_metadata>
{data1}
</ground_truth_metadata>

Generated Caption Metadata (DATA2):
<generated_caption_metadata>
{data2}
</generated_caption_metadata>

Your task is to compare these two sets of metadata and rate their similarity on a scale of 1 to 10, where 1 indicates completely different speaking-styles and 10 indicates identical speaking-styles.

Instructions:
1. Analyze and compare the following attributes between DATA1 and DATA2:
  a) Speaker gender
  b) Accent
  c) Pitch
  d) Speaking rate
  e) Environment/noise level
  f) Emotion/expression
  g) Voice quality (e.g., guttural, silky)
  h) Prosody (including rhythm, intonation, and stress)
  i) Intention of speech

2. Inside your thinking block, wrap your attribute comparisons inside <attribute_comparison> tags. For each attribute:
  - Rate its similarity on a scale of 1-10
  - Summarize key similarities and differences
  - Consider how they contribute to the overall speaking-style
  Pay special attention to stylistic aspects, intention, and prosody.

3. After analyzing all attributes, calculate an average similarity score based on your individual attribute ratings.

4. Don't penalize the generation if is more specific than the ground truth. For example, if the ground truth says "American accent" and the generation says "Boston accent", that's fine.

5. Provide a comprehensive justification for your similarity rating in <justification> tags. Explain how the similarities and differences in each attribute contribute to your overall assessment.

6. In the metadata, only using the following information:
  - speaker gender
  - accent
  - pitch
  - speaking rate
  - environment/noise level
  - emotion/expression

7. Finally, provide your similarity score in JSON format within <score> tags. The JSON should include a single key 'similarity_score' with a value between 1 and 10, based on your calculated average and overall assessment.

Example output structure (do not copy the content, only the structure):

<attribute_comparison>
a) Speaker gender:
  Similarity rating: X/10
  Key similarities: [...]
  Key differences: [...]
  Contribution to overall style: [...]
b) Accent:
  Similarity rating: X/10
  Key similarities: [...]
  Key differences: [...]
  Contribution to overall style: [...]
[... continue for all attributes ...]
Average similarity score: X/10
</attribute_comparison>

<justification>
[Your comprehensive justification for the similarity rating, synthesizing the analyses of individual attributes]
</justification>

<score>
{{'similarity_score': X}}
</score>

Remember to focus on the stylistic aspects of speech, paying particular attention to the intention behind the speech and the nuances of prosody in your analysis and justification.

Your final output should consist only of the justification and score, without duplicating the detailed attribute comparisons from your thinking block.
```

Figure 9: Prompt template for rating speaking-style similarity between ground-truth metadata and generated-caption metadata.

```
You are an expert evaluator assessing the quality of automatically generated audio captions. Your task is to compare metadata from audio files with generated captions to identify and score hallucinations.
## Task Overview
Compare the **metadata** (ground truth information about the audio) with a **generated caption** and evaluate the presence and severity of hallucinations.
## Input Format
Ground Truth Metadata:
<ground_truth_metadata>
{data1}
</ground_truth_metadata>
Generated Caption:
<generated_caption>
{data2}
</generated_caption>
## Evaluation Criteria
### What Counts as a Hallucination?
A hallucination is information in the generated caption that:
- Directly contradicts the metadata
- Invents specific details not supported by any evidence
- Misrepresents factual characteristics (e.g., wrong gender, wrong emotion, wrong accent)
### What Does NOT Count as a Hallucination?
The following are **acceptable creative interpretations** and should NOT be penalized:
- **Reasonable emotional inferences**: If metadata shows "laughing" audio, describing the speaker as "cheerful," "animated," or "enthusiastic" is acceptable
- **Stylistic embellishments**: Descriptive language like "her voice carries warmth," "delivered with precision," or "words tumble out" that enhance the description without contradicting facts
- **Contextual scenarios**: Adding reasonable context (e.g., "as if addressing an audience," "in a quiet environment") when supported by acoustic characteristics
- **Prosodic interpretations**: Descriptions of rhythm, cadence, emphasis, or delivery style that are reasonable given the speech characteristics
- **Minor elaborations**: Details that naturally extend from the metadata without contradicting it
### Hallucination Examples
**SEVERE Hallucinations:**
- Metadata indicates male speaker -> Caption says "female speaker"
- Metadata shows speaker p026 -> Caption invents specific name or identity
- Metadata indicates disgust emotion -> Caption describes joy or happiness
- Inventing specific events or quoted speech not in the audio
**MODERATE Hallucinations:**
- Claiming "noisy environment" when metadata suggests clean audio
- Describing specific background sounds not mentioned in metadata
- Asserting definitive context (e.g., "delivering a news broadcast") without support
**MINOR Hallucinations:**
- Slightly exaggerating a characteristic (e.g., "booming voice" for a moderately loud voice)
- Adding plausible but unverified details that don't contradict metadata
## Scoring Instructions
Provide two scores:
### 1. Hallucination Count (0-10 scale)
- **0-2**: No hallucinations or only very minor embellishments
- **3-4**: 1-2 minor hallucinations that don't significantly distort the audio
- **5-6**: Multiple minor hallucinations OR 1 moderate hallucination
- **7-8**: Multiple moderate hallucinations OR 1 severe hallucination
- **9-10**: Multiple severe hallucinations that fundamentally misrepresent the audio
### 2. Severity Level
- **Low**: Only minor embellishments; caption is largely accurate
- **Medium**: Some factual errors that partially misrepresent the audio
- **High**: Severe contradictions or fabrications that fundamentally mischaracterize the audio
## Instructions
1. In your thinking block, carefully analyze the metadata and generated caption step by step:
   - List all factual claims made in the generated caption
   - For each claim, check if it's supported by, contradicts, or extends beyond the metadata
   - Categorize each issue as minor, moderate, or severe
   - Count the total number of hallucinations
2. Be lenient with creative, descriptive language that doesn't contradict facts
3. Focus on factual accuracy regarding speaker identity, emotions, accent, and acoustic characteristics
4. Distinguish between reasonable inference and baseless fabrication
## Output Format
After your thinking, provide your evaluation in the following structure:
<analysis>
[Brief explanation of what you observed, noting any hallucinations found and why they are/aren't problematic]
</analysis>
<justification>
Explain your scoring in detail, referencing specific examples from the caption and how they relate to the metadata
</justification>
<score>
{{"hallucination_count": X, "severity": "Low/Medium/High"}}
</score>
Remember: Context is important - "laughing" audio naturally suggests positive emotions. Be strict about factual contradictions but lenient about creative descriptions that align with the metadata.
Your final output should consist only of the analysis, justification and score, without duplicating the detailed step-by-step analysis from your thinking block.
```

Figure 10: Hallucination evaluation prompt for comparing ground-truth metadata against a generated caption.

```
You are an expert evaluator specializing in assessing instruction-following capabilities of audio-language models,
particularly in the domain of speaking-style captioning. Your role is to rigorously evaluate whether generated captions
accurately follow given instructions while maintaining factual accuracy about the audio content.
## Context
audio-language models are tasked with generating descriptive captions about speaking-styles, vocal characteristics, and
acoustic properties of speech recordings. These models receive specific instructions that guide what aspects to focus on,
what style to adopt, or what format to use in their captions.
## Your Task
Evaluate how well the generated caption follows the provided instruction, considering both explicit requirements and implicit
expectations.
—
## Input Materials
### Instruction Given to the Model:
<instruction>
{instruction}
</instruction>
### Generated Caption by the Model:
<generated_caption>
{caption}
</generated_caption>
—
## Evaluation Framework
### 1. **Instruction Compliance Categories**
Analyze the instruction for the following types of requirements:
**A. Content Requirements**
- Specific attributes to include (e.g., "describe the pitch," "mention the accent")
- Attributes to exclude or de-emphasize
- Level of detail requested (brief, detailed, comprehensive)
- Factual accuracy requirements
**B. Structural Requirements**
- Format specifications (paragraph, bullet points, structured description)
- Length constraints (word count, sentence count)
- Organizational pattern (chronological, importance-based, categorical)
**C. Stylistic Requirements**
- Tone (formal, casual, technical, creative)
- Perspective (objective observer, subjective interpreter, technical analyst)
- Language complexity (simple, sophisticated, jargon-heavy)
- Writing style (descriptive, analytical, narrative)
**D. Focus Requirements**
- Specific aspects to emphasize (prosody, emotion, acoustic features)
- Target audience considerations
- Use case or application context
**E. Constraint Requirements**
- Prohibitions (avoid certain terms, don't speculate)
- Boundaries (stick to observable features only)
- Scope limitations (focus on X, ignore Y)
### 2. **Evaluation Criteria**
For each requirement identified in the instruction, assess:
**Explicit Compliance (70% weight)**
- Does the caption directly address explicitly stated requirements?
- Are all mandatory elements present?
- Are any prohibited elements absent?
- Is the specified format/structure followed?
**Implicit Compliance (20% weight)**
- Does the caption follow reasonable interpretations of the instruction?
- Are contextual clues and implied expectations honored?
- Does it align with the spirit/intent of the instruction?
**Quality of Execution (10% weight)**
- How well are the requirements implemented (not just their presence)?
- Is the execution natural and coherent, or forced and awkward?
- Does following the instruction enhance or detract from caption quality?
### 3. **Common Instruction Types in Audio Captioning**
Be aware of these typical instruction patterns:
- **Attribute-focused**: "Describe the speaker's pitch and speaking rate"
- **Style-focused**: "Write in a technical, objective manner"
- **Audience-focused**: "Explain as if to a non-expert"
- **Format-focused**: "Provide a structured analysis with separate sections"
- **Perspective-focused**: "Focus on perceptual qualities rather than technical measurements"
- **Comparative**: "Compare this to typical conversational speech"
- **Holistic vs. Analytical**: "Provide an overall impression" vs. "Break down individual components"
```

Figure 11: Instruction-following evaluation prompt (Part 1): context, inputs, and evaluation framework overview.

```
—
## Scoring Guidelines
### Instruction-Following Success: Success / Partial Success / Failure
**Success**:
- All critical requirements met (90%+ compliance)
- No significant violations of explicit constraints
- Natural execution that maintains caption quality
- Minor omissions only in non-essential elements
**Partial Success**:
- Most requirements met (60-89% compliance)
- Some explicit requirements missed or violated
- Execution may be somewhat awkward or forced
- Core intent of instruction honored despite gaps
**Failure**:
- Major requirements ignored (<60% compliance)
- Explicit constraints violated
- Instruction intent misunderstood or disregarded
- Caption appears written without considering the instruction
### Instruction-Following Accuracy: 1–10 Scale
**9–10 (Exceptional)**: Near-perfect instruction adherence. All explicit and implicit requirements met with natural,
high-quality execution. Any deviations are trivial.
**7–8 (Strong)**: Excellent instruction-following with minor gaps. All critical requirements met. May miss 1–2 minor
elements or execute some aspects imperfectly, but overall very aligned.
**5–6 (Adequate)**: Satisfactory instruction-following with notable gaps. Core requirements met but several secondary
elements missed or poorly executed. Instruction intent generally honored.
**3–4 (Weak)**: Poor instruction-following with major gaps. Many requirements missed or violated. Some alignment with
instruction but significant divergence in execution.
**1–2 (Very Weak)**: Minimal instruction-following. Most requirements ignored. Caption appears largely independent of the
instruction. Major misunderstanding or disregard of requirements.
—
## Special Considerations
### Edge Cases to Watch For:
1. **Conflicting Requirements**: If instruction contains contradictory elements, evaluate based on how well the model
navigates the conflict.
2. **Ambiguous Instructions**: If instruction is vague, give credit for reasonable interpretations that align with common
sense.
3. **Impossible Requirements**: If instruction asks for information not determinable from audio alone, don't penalize the
model for not hallucinating content.
4. **Over-compliance**: If model follows instruction so rigidly that caption quality suffers dramatically, note this in
justification but still credit the instruction-following attempt.
5. **Creative Interpretation**: Distinguish between helpful creative interpretation (good) and ignoring instructions (bad).
### Common Pitfalls to Identify:
- ✘ **Instruction ignored entirely**: Caption is generic and could apply to any instruction
- ✘ **Partial attention**: Model addresses only the first part of a multi-part instruction
- ✘ **Format violation**: Instruction specifies structure but caption uses different format
- ✘ **Tone mismatch**: Formal instruction produces casual output, or vice versa
- ✘ **Scope creep**: Model includes content explicitly excluded by instruction
- ✘ **Misinterpretation**: Model misunderstands key terms or intent of instruction
—
## Output Format
Provide your evaluation in the following structure:
<analysis> [Identify the key requirements in the instruction: What are the explicit demands? What are the implicit
expectations? What constraints are specified? Categorize them as content, structural, stylistic, focus, or constraint
requirements.] </analysis>
<compliance_assessment> [For each identified requirement, evaluate whether the caption meets it. Mark each as: ✓ Met, △
Partially Met, or ✗ Not Met. Provide brief evidence from the caption.] </compliance_assessment>
<strengths> [Highlight what the model did well in following the instruction. What requirements were executed particularly
effectively?] </strengths>
<weaknesses> [Identify gaps, violations, or missed opportunities. What requirements were ignored, poorly executed, or
misinterpreted?] </weaknesses>
<justification> [Synthesize your analysis into a coherent evaluation. Explain your success determination and accuracy score.
Reference specific examples from both the instruction and caption. Consider the severity of any violations and the overall
alignment with instruction intent.] </justification>
<ratings>
{ "instruction_following_success": "[Success/Partial Success/Failure]", "instruction_following_accuracy": [1-10] }
</ratings>
—
## Important Reminders
1. **Focus on instruction compliance, not caption quality**: A poorly written caption that follows instructions perfectly
should score higher than a beautifully written caption that ignores them.
2. **Be precise in your reasoning**: Cite specific phrases from both the instruction and caption to support your evaluation.
3. **Distinguish between critical and minor requirements**: Not all instruction elements carry equal weight. Prioritize
core requirements over secondary details.
4. **Consider context**: Audio captioning has domain-specific norms. Evaluate within the context of speaking-style
description conventions.
5. **Be fair but rigorous**: Give credit for good-faith attempts while maintaining high standards for what constitutes
successful instruction-following.
6. **Avoid lenience bias**: Don't give full credit just because the caption is reasonable. It must specifically follow the
given instruction.
Begin your evaluation now.
```

Figure 12: Instruction-following evaluation prompt (Part 2): scoring rubric, edge cases, and required output format.

```
{
"unseen_prompt_1": "Establish the speaker's vocal identity without mentioning or inferring gender. Use only acoustic cues to describe pitch range, timbre/texture, resonance, and any stable voice qualities; if a trait is uncertain, state it as unknown rather than guessing.",
"unseen_prompt_2": "Write a speaker-identity caption that avoids any accent or origin claims. Focus on voice quality (e.g., breathiness/creak/clarity), pitch behavior, articulation style, and timbral color, relying strictly on what can be heard.",
"unseen_prompt_3": "Output a Python-dict-style profile (not JSON) with exactly these keys: {p̈itch_character̈: ..., ẗimbrë: ..., v̈ocal_texturë: ..., s̈tabilitÿ: ..., n̈otes̈: ...}. Fill values with short phrases grounded in the audio; use null for anything you cannot justify.",
"unseen_prompt_4": "Provide two competing hypotheses for the speaker's vocal identity (Hypothesis A vs B) based only on acoustics, then pick the better one with a confidence score from 0 to 1. Do not mention language, name, or demographics.",
"unseen_prompt_5": "Describe the speaker as a casting director evaluating a voice actor for a role. Keep it strictly evidence-based: every descriptive claim must correspond to an audible cue (pitch movement, timbre, articulation, resonance).",
"unseen_prompt_6": "Produce a single sentence that captures the speaker's 'voiceprint' using neutral phrasing and no demographic attributes. The sentence must include exactly one timbre descriptor, one pitch descriptor, and one articulation descriptor.",

  "unseen_prompt_7": "Infer the recording context (environment + microphone distance) and present it as: (1) best-guess setting, (2) 2-3 acoustic cues supporting it, (3) an uncertainty note. Do not mention emotion or speaker identity.",
"unseen_prompt_8": "Write a 'field recording log' in three bullet points: Background (noise type + stability), Channel (reverb/roominess + clipping/cleanliness), Distance (near/far + evidence). Avoid guessing location names.",
"unseen_prompt_9": "Return a compact YAML block with keys: setting_guess, noise_profile, reverberation, mic_distance, confidence. Values must be short, and confidence must be one of: low/medium/high.",
"unseen_prompt_10": "Describe how the environment changes over time across the utterance (start → middle → end). Only reference audible changes (noise bursts, reverb shifts, level changes); do not hallucinate events.",
"unseen_prompt_11": "As an audio engineer, diagnose whether the speech sounds studio-like or in-the-wild and justify your conclusion with at least three acoustic indicators. If evidence is mixed, explicitly say so.",
"unseen_prompt_12": "Give a one-sentence context caption that must avoid these words: room, background, noise, microphone. You may still describe context using alternative phrasing grounded in acoustics.",

  "unseen_prompt_13": "Describe the affect using only valence (1-7) and arousal (1-7), plus short acoustic evidence (pitch dynamics, intensity, tempo, voice quality). Do not use categorical emotion labels.",
"unseen_prompt_14": "Return a strict JSON object with exactly these keys: {v̈alencë:1-7, ärousal̈:1-7, c̈onfidencë:0-1, ëvidencë:[...]}.",
"unseen_prompt_15": "Write two sentences: Sentence 1 must be purely objective (acoustic description). Sentence 2 may interpret the likely feeling state but must include a caution clause about uncertainty.",
"unseen_prompt_16": "Describe the emotional trajectory across the utterance and name the specific prosodic features that signal each phase. Avoid any mention of the textual content.",
"unseen_prompt_17": "As a clinician writing a neutral intake note, summarize the speaker's affective presentation using cautious language and explicit acoustic grounding; do not overclaim.",
"unseen_prompt_18": "Generate a caption that includes exactly one metaphor for the affect, immediately followed by concrete acoustic justification.",

  "unseen_prompt_19": "Infer the likely communicative intent from prosody alone and provide 2-3 prosodic cues as justification.",
"unseen_prompt_20": "Output a ranked list of up to three pragmatic functions with probability estimates that sum to 1.0.",
"unseen_prompt_21": "Write as a debate judge assessing delivery: identify confidence vs hesitation and strategic emphasis.",
"unseen_prompt_22": "Produce a minimal-pair analysis contrasting the most likely intent with a plausible alternative.",
"unseen_prompt_23": "Return a markdown table: Prosodic cue | What it signals | Confidence (low/med/high).",
"unseen_prompt_24": "Create a single-sentence pragmatic caption that must not include any pitch words.",

  "unseen_prompt_25": "Rate warmth, urgency, and confidence (1-5) with one acoustic justification per scale.",
"unseen_prompt_26": "As a call-center QA evaluator, describe perceived clarity and assertiveness, separating observation from interpretation.",
"unseen_prompt_27": "Write a listener-centric caption focusing only on intelligibility and comfort.",
"unseen_prompt_28": "Return a strict JSON object with keys: impression, accessibility_risks, most_salient_cue, confidence.",
"unseen_prompt_29": "Advise a TTS system how to match the perceived presence and approachability using audible cues only.",
"unseen_prompt_30": "Give two captions: one for a casual listener and one for a trained listener with technical descriptors.",

  "unseen_prompt_31": "Write director's notes for recreating this performance, grounded strictly in acoustics.",
"unseen_prompt_32": "Produce two labeled lines: FACT and IMAGERY, ensuring IMAGERY does not contradict FACT.",
"unseen_prompt_33": "Return a nested JSON object summarizing speaker, delivery, and context; use null for unknowns.",
"unseen_prompt_34": "Write a sports-commentator-style description justified by explicit acoustic cues.",
"unseen_prompt_35": "Create a single sentence of exactly 18 words integrating identity, delivery, and context.",
"unseen_prompt_36": "Write a three-bullet spec sheet (Identity / Delivery / Context) with acoustic justification."
}
```

Figure 13: Unseen instruction prompts used to probe diverse instruction-following behaviors for speaking-style captioning.

I Model Training Details

Models	Role	Sources
Audio-Flamingo-3	Evaluated	https://huggingface.co/nvidia/audio-flamingo-3
Voxtral-Mini-3B	Evaluated	https://huggingface.co/mistralai/Voxtral-Mini-3B-2507
Voxtral-Small-24B	Evaluated	https://huggingface.co/mistralai/Voxtral-Small-24B-2507
Qwen2-Audio-7B-Instruct	Evaluated	https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct
MERaLiON-AudioLLM-Whisper-SEA-LION	Evaluated	https://huggingface.co/MERaLiON/MERaLiON-AudioLLM-Whisper-SEA-LION
Qwen3-Omni	Evaluated	https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Instruct
MOSS-Audio	Evaluated	https://github.com/OpenMOSS/MOSS-Audio
Audio-Flamingo-Next	Evaluated	https://huggingface.co/nvidia/audio-flamingo-next-hf
SALMONN-7B	Evaluated	https://github.com/bytedance/SALMONN/tree/salmonn
OpenAI-GPT-Audio	Evaluated (closed-source reference)	https://platform.openai.com/docs/models/gpt-audio
Mistral-Small-24B-Instruct	Caption generation for train/validation	https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501
Llama-3.1-8B-Instruct	Caption generation for evaluation splits	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
Gemma-2-27B	Caption generation for adversarial caption set	https://huggingface.co/google/gemma-2-27b
Qwen-3-32B	LLM-Judge	https://huggingface.co/Qwen/Qwen3-32B
Gemma-4-31B-it	LLM-Judge	https://huggingface.co/google/gemma-4-31B-it
GPT-OSS-20B	LLM-Judge	https://huggingface.co/openai/gpt-oss-20b

Table 19: Specification and sources of the employed models.

Training and inference details. We fine-tune the SALMONN variants using LoRA to control adaptation capacity and limit task-specific collapse. We set the LoRA rank to r=8 throughout. During training, we use a stronger scaling (α=32) to enable rapid alignment to the instruction-following format and the speech-conditioning interface; at inference, we reduce the scaling to α=16. This inference-time attenuation acts as a practical "de-overfitting" knob: it dampens overly task-specialized update directions while preserving the base LLM's general generation ability, improving robustness to prompt and domain shifts. Empirically, this reduces random stylistic drift (e.g., ungrounded claims about affect or speaker traits) and encourages generations to remain anchored to the acoustic evidence rather than the most frequent training-task patterns.

Instruction-tuning regimes and batch mixing. We study two instruction-tuning regimes. **Single-task** tuning introduces one new task, which isolates the effect of task supervision but also makes memorization easier. **Multi-task** tuning introduces six new tasks, where tasks are sampled uniformly at random within each mini-batch. This stochastic task mixture increases gradient diversity and discourages the model from specializing to any one instruction template. We additionally explored structured batch sampling strategies (e.g., round-robin task cycling and scheduled curricula), but these did not improve performance and in several cases degraded it, likely because deterministic mixing reduces the regularizing effect of interleaving and amplifies short-horizon task dominance within contiguous training windows.

Scaling data and diagnosing task overfitting. Training on the PSC-base human-annotated subset (282 hours) yields a consistent failure mode: (i) the model overfits to the limited set of training tasks, (ii) it produces less expressive captions (tending toward safe, generic phrasing), and (iii) it exhibits catastrophic forgetting, struggling to answer open-set questions that were not explicitly present in the tuning set. We interpret this as a distributional "compression" effect: with limited diversity, the model learns a narrow mapping from acoustics to a small set of instruction-conditioned outputs, trading off nuanced style description for high-confidence task compliance. Increasing scale mitigates these issues: training on 100% of PSC and expanding instruction diversity to roughly $\sim$300 prompts per task substantially improves caption quality and stylistic richness. However, the underlying task-overfitting tendency persists, suggesting that capacity control and objective design (not only data volume) are necessary to fully address specialization.

Alternative learning strategies for style grounding. To improve grounding and reduce spurious style attributions, we investigate two complementary strategies. (**A**) **Auxiliary-task inclusion:** we incorporate automatic speech recognition (ASR), gender recognition (GR), and accent recognition (AR) as additional supervised objectives. These tasks encourage the shared representation to encode linguistically and paralinguistically salient cues, acting as an inductive bias toward acoustically-verifiable factors. (**B**) **Factor-conditioned captioning:** we modify the training target to first predict explicit factors (e.g., `gender`, `accent`, `pitch`, `speaking rate`) followed by the free-form speaking-style caption. By forcing the model to commit to concrete attributes before generating the descriptive caption, we aim to reduce shortcut learning and improve causal alignment between acoustics and the final narrative description.

Transcript as linguistic context. We provide the transcript as linguistic context for speaking-style captioning. This yields a complementary perspective: it allows the model to separate *what* was said (lexical content) from *how* it was said (prosody, voice quality, pacing), and to interpret paralinguistic signals relative to discourse intent (e.g., emphasis, question intonation, sarcasm markers). To our knowledge, explicitly incorporating the transcript in this way has not been standard in prior speaking-style captioning setups, and it improves the model's ability to generate contextually appropriate style descriptions rather than purely acoustic correlates.

Decoding. At inference time, we use default decoding settings for all models unless otherwise noted. For SALMONN, we use deterministic beam search to prioritize stability and reduce sampling-induced variance: `max_new_tokens=200`, `num_beams=4`, `do_sample=False`, `min_length=1`, `temperature=1.0`, `top_p=0.9`, `repetition_penalty=1.0`, `length_penalty=1.0`.

J Lexical Diversity

From the lexical diversity analysis in Table 20, we find that task-specific fine-tuning generally reduces linguistic diversity, which is expected since the model is optimized to consistently express a constrained set of speaking-style attributes rather than maximize unconstrained lexical variation. Compared with single-task training, multi-task training improves faithfulness and reduces hallucinations (Table 1), but its effect on diversity is mixed: it clearly reduces repetition across captions (Self-BLEU 0.206 vs. 0.315), whereas HD-D and Distinct-3 are slightly lower and Distinct-2 is essentially unchanged. A larger LLM backbone also does not consistently increase diversity. From Voxtral-Mini-3B to Voxtral-Small-24B, Distinct-2 and Distinct-3 increase, but HD-D decreases slightly and Self-BLEU rises, indicating more repetition. We additionally observe that the oracle captions in StyleInstructCaps are considerably more diverse than those in ParaSpeechCaps. Models trained on StyleInstructCaps are not uniformly more diverse than the caption set itself. Both SALMONN (SIC-DB) variants exceed the oracle captions on HD-D, and the MT variant has lower Self-BLEU, but both have lower Distinct-2 and Distinct-3.

Caption set	HD-D	Self-BLEU	Distinct-2	Distinct-3	Mean caption length	STD of caption lengths
Oracle caption (PSC)	0.732	—	0.0425	0.0795	32.5	7.97
Oracle caption (SIC)	0.771	0.304	0.152	0.310	34.9	12.5
Audio-Flamingo-3	0.748	0.0881	0.138	0.248	31.8	23.0
Voxtral-Mini-3B	0.813	0.123	0.247	0.463	145	78.1
Voxtral-Small-24B	0.807	0.145	0.281	0.484	72.0	56.8
Qwen2-Audio-7B-Instruct	0.836	0.0362	0.231	0.439	86.2	59.5
MERaLiON-AudioLLM	0.782	0.0826	0.259	0.429	29.9	21.6
MOSS-Audio	0.836	0.0478	0.342	0.631	56.6	26.4
Qwen3-omni	0.856	0.0641	0.343	0.610	134	92.0
Audio-Flamingo-next	0.885	0.128	0.330	0.501	46.8	44.4
SALMONN	0.795	0.113	0.221	0.397	84.3	32.4
SALMONN (PSC)	0.735	0.361	0.0906	0.195	66.4	18.9
SALMONN (SIC-DB) - ST	0.786	0.315	0.135	0.271	72.8	15.4
SALMONN (SIC-DB) - MT	0.776	0.206	0.137	0.266	52.8	18.0

Table 20: Lexical diversity and caption-length statistics across different caption sets and models.

K LLM Judgement

LLM judgement

similarity_score: 7

justification:

The generated caption matches the ground truth in gender, accent, pitch, and speaking rate. It captures the emotional tone of discomfort and frustration, though it's slightly less specific than the ground truth's pained emotion. The generated caption introduces a nasal voice quality not mentioned in the ground truth, which is acceptable as per instructions. The environment/noise level is missing, which is a notable difference.

Ground truth metadata

{ "source": "ears", "relative_audio_path": "p019/emo_pain_freeform.wav", "text_description": [" A male speaker with an American accent delivers a speech that is deep and high-pitched, characterized by crisp, husky tones. His voice is measured in speed, punctuated with slight pauses, conveying a sense of pained emotion. The speech was recorded in a slightly noisy environment, adding a flowing, textured quality to the overall sound.", " A male speaker with a high-pitched voice delivers pained expressions in an American accent, speaking at a measured speed in a slightly noisy environment."], "transcription": " Ow, god damn it, that hurts. Damn, I wish I hadn't stubbed my toe on the corner of that table. Oh, that's kind of smart. Oh, boy, I hate that so much. Why did it have to be the pinky toe as well? Ow.", "intrinsic_tags": ["crisp", "deep", "american", "flowing", "husky", "punctuated"], "situational_tags": ["pained"], "basic_tags": ["high-pitched", "male", "measured speed", "slightly noisy environment"], "all_tags": ["american", "crisp", "deep", "flowing", "high-pitched", "husky", "male", "measured speed", "pained", "punctuated", "slightly noisy environment"], "speakerid": "p019", "name": "p019", "duration": 16.1, "gender": "male", "accent": "american", "pitch": "high-pitched", "speaking_rate": "measured speed", "noise": "slightly noisy environment", "utterance_pitch_mean": 159.45840454101562, "snr": 38.82289123535156, "holistic_creative_synthesis": "A high-pitched, crisp, and husky American male voice speaks in a slightly noisy environment, conveying a sense of pained emotion through a measured pace and deliberate pauses, the speaker's words punctuated by a rhythmic flow that adds to the overall sense of conviction, creating a sense of depth and nuance that draws the listener in.", "holistic_creative_synthesis_instruction": "Combine all elements of the speaker's style, including their accent, pitch, timbre, and delivery, to create a rich, narrative-style caption that captures the essence of their speaking-style." }

Generated caption

Generated caption: The speaker's style is characterized by a low, guttural tone, a slow and deliberate pace, and a deep, resonant timbre. Their accent is distinct and evocative, with a strong emphasis on certain syllables and a distinctive rhythm. The delivery is confident and assertive, with a sense of authority and expertise. Overall, the speaker's style is engaging and captivating, drawing the listener in with their unique voice and delivery.

Figure 14: Example showing an LLM judgment (top) and the corresponding ground-truth metadata vs. generated caption (bottom).

L Speaker Caption clustering plots

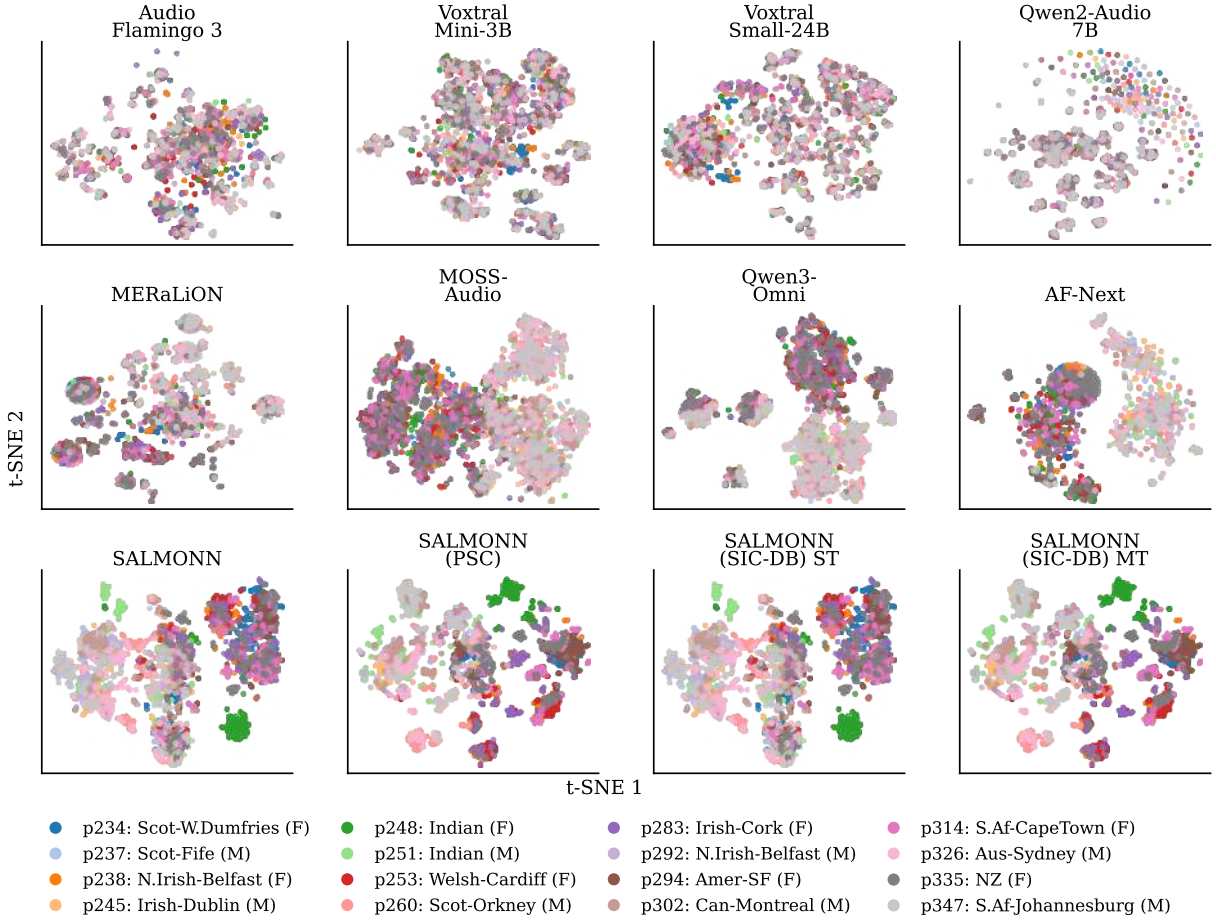


Figure 15: t-SNE visualization of speaker-style caption embeddings on UD-VCTK (all speakers)

M Use of AI assistants

We used GPT-5 and Gemini 3 Flash for code assistance, and the same models purely for language-related assistance in writing the paper.

N Licensing

All resources used in this work are released under open licenses. We use the VCTK corpus, which is distributed under the Open Data Commons Attribution License (ODC-By) v1.0, requiring attribution. We also use the GigaSpeech dataset, released under the Apache License 2.0. ParaSpeechCaps is provided under the CC BY-NC-SA 4.0 license, which restricts commercial use and requires attribution and share-alike for derivatives. Finally, the SALMONN codebase is released under the Apache License 2.0.