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

Anonymous ACL submission

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 task formulation that supports controllability over which stylistic dimensions should be described. SSC spans abstraction levels ranging from low-level acoustic traits to broad, context-dependent characterizations, making comparison and evaluation difficult. 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. Based on this formulation, we introduce *StyleInstructCaps*, the first standardized benchmark for controllable speaking-style captioning in audio-language models. *StyleInstructCaps* provides a task-specific dataset and an evaluation framework that measures metadata groundedness, hallucination, instruction-following ability, speaker style consistency, and generalization to unseen instructions and audio datasets.[1]

1 Introduction

Speaking-style captioning (SSC) has recently emerged as a promising direction for automating the description of how speech is delivered, capturing properties such as vocal timbre, prosody, rhythm, emphasis, emotion and other paralinguistic cues in natural language (Yamauchi et al., 2024). Yet, the field still lacks a shared understanding of what the task represents.

[1]We will release the code, data, and benchmark upon acceptance.

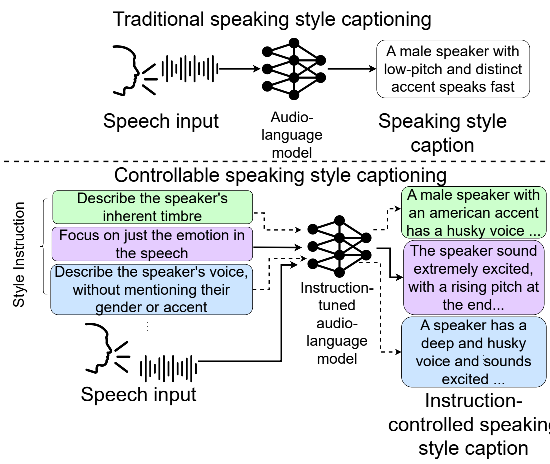


Figure 1: Speaking-style captioning as an instruction-following audio-language task

Researchers in audio understanding often view SSC as an extension of speech comprehension, with an emphasis on interpreting communicative intent and discourse-level meaning (Jing et al., 2024; Chandra et al., 2025; Zhang et al., 2025b). Researchers in speech synthesis and generative modeling, in contrast, tend to treat speaking-style captions as descriptions of vocal quality that can be used to control expressive text-to-speech systems (Wang et al., 2025a; Guo et al., 2023) and voice conversion models (Kuan et al., 2023; Hai et al., 2024). These differences reflect a deeper challenge. **The notion of speaking-style is inherently subjective** (Adigwe et al., 2024), and different research communities anchor the notion of style in different aspects of the speech signal. As a result, it becomes difficult to compare models fairly, interpret results consistently, and measure performance across studies uniformly.

To mitigate this discrepancy, 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 approaches treat SSC as a single task that produces a holistic description of speech, often entangling speaker traits, emotion, and delivery characteristics, which makes 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. This formulation supports diverse downstream applications, including expressive text to speech, voice conversion, conversational agents, and large scale speech annotation, and improves speech understanding by enabling models to attend selectively to relevant paralinguistic cues rather than producing vague or conflated descriptions of speaking-style.

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) **speaker-style consistency**, and (5) **generalization to unseen speech data and prompts**.

To support this benchmark, we construct a multi-style speaking-style caption corpus built on top of the ParaSpeechCaps (PSC) dataset (Diwan et al., 2025a). The dataset contains 2.9k hours of speech, and each utterance is paired with six complementary caption types that target distinct dimensions of speaking-style: speaker idiosyncratic style, situational/contextual delivery style, expressive/emotion style, linguistic pragmatic style, perceptual listener-centric style, and an overall style summary. These parallel annotations enable controlled evaluation of instruction-following and allow us to probe whether models can isolate individual style dimensions, integrate multiple cues when required, and generalize across diverse speaking conditions.

We further conduct a systematic comparison of six state-of-the-art audio-language models: Audio-Flamingo-3 (Goel et al., 2025), Voxtral-Mini-3B (Liu et al., 2025), Voxtral-Small-24B (Liu et al., 2025), Qwen2-Audio-7B-Instruct (Chu et al., 2024), MERaLiON (MERaLiON Team, 2024), and SALMONN (Tang et al., 2024). In particular, we present detailed experiments with SALMONN for two practical reasons. First, SALMONN is a widely used and fully open audio-language model with a modular encoder–LLM design, making it feasible to run controlled ablations and instruction-tuning variants while holding architecture constant. Second, its training recipe and speech encoders are strongly aligned with paralinguistic understanding, providing a representative and reproducible testbed for isolating the effects of (1) SSC-specific supervision and (2) instruction diversity in controllable SSC. Through this work, we aim to establish a unified and rigorous foundation for research on controllable SSC, provide standardized resources for reproducible evaluation, and enable progress toward audio-language models that produce faithful, interpretable, and instruction-aware descriptions of how people speak.

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 2.9k-hour multi-style speaking-style caption corpus with six parallel caption types per utterance; (3) We evaluate six state-of-the-art audio-language models, **establishing strong baselines** and revealing substantial room for improvement on controllable SSC.

2 Related Work

2.1 Speaking-Style Captioning for Paralinguistic Speech Understanding

Speaking-style captioning (SSC) aims to generate natural-language descriptions of *how* speech is delivered, capturing paralinguistic attributes such as vocal timbre, prosody, emotion, and delivery style (Yamauchi et al., 2024). Compared to categorical paralinguistic tasks such as speaker recognition (Hansen and Hasan, 2015), speech emotion recognition (Triantafyllopoulos et al., 2025), speech question answering (Peng et al., 2024), and speech summarization (Shang et al., 2024; Retkowski et al., 2025), SSC produces open-ended descriptions that can integrate multiple perceptual cues into a unified account. This open-ended interface has become increasingly relevant in the era of large audio-language models, where natural-language explanations and descriptions serve as a flexible, interpretable layer for expressing subjective speaking-style phenomena (Su et al., 2025; Dhamyal et al., 2024; Lu et al., 2024, 2025).

2.2 Datasets and Benchmarks

Progress in SSC depends on (1) speech corpora with rich style-related supervision and (2) evaluation protocols that can assess generation quality beyond lexical overlap. Existing resources span multiple communities: expressive TTS and style-prompted speech datasets provide fine-grained speaker and delivery attributes (e.g., prompt- or caption-supervised corpora), while paralinguistic understanding benchmarks evaluate reasoning over non-verbal cues such as emotion, prosody, and conversational pragmatics. Appendix A Table 4 and Table 5 summarize representative datasets and benchmarks in this space. However, most existing benchmarks emphasize *recognition* rather than SSC as a standalone *generation* task, and they typically do not support *controllable* evaluation over which stylistic dimension should be described (e.g., isolating speaker traits vs. emotion vs. contextual delivery). This mismatch makes it difficult to compare SSC models consistently and to test whether a model can selectively attend to the 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., 2025) 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., 2022a) 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

Motivated by the gaps between SSC task and existing benchmarks, we build on ParaSpeechCaps (Diwan et al., 2025a), which provides large-scale speech paired with per-utterance metadata covering both intrinsic speaker characteristics and situational delivery attributes, and we introduce a controllable SSC benchmark where instructions explicitly specify the targeted speaking-style dimensions.

3.1 Problem Formulation

We introduce *controllable SSC*, a conditional generation task in which a model produces natural language descriptions of how speech is delivered. Given a speech signal and an explicit instruction specifying a target stylistic dimension, such as speaker-specific traits, emotional expression, or delivery style, the model generates a caption describing the requested aspect of speaking-style. Unlike traditional SSC, which produces a single holistic description per utterance, controllable SSC enables targeted, instruction-conditioned descriptions that isolate specific stylistic dimensions. This formulation supports systematic evaluation and better aligns SSC with downstream speech understanding and generation applications. To unify these training configurations under a common learning framework, we formalize the instruction-tuning objective used for controllable SSC below.

Let $\mathcal{D} = \{(x, s, y)\}$, where x is the input, s is the style instruction, and $y = (y_1, \ldots, y_T)$ is the target speaking-style caption. We instruction-tune by minimizing the negative log-likelihood conditioned on (x, s):

$$\begin{aligned}\mathcal{L}_{\text{SFT}}(\theta) &= \mathbb{E}_{(x,s,y)\sim\mathcal{D}}[-\log p_\theta(y \mid x, s)] \\ &= \mathbb{E}_{(x,s,y)\sim\mathcal{D}}\left[\sum_{t=1}^{T} -\log p_\theta(y_t \mid x, s, y_{<t})\right].\end{aligned} \tag{255}$$

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., 2025a). PSC aggregates multiple speech datasets and comprises two subsets: PSC-base, which includes Expresso (Nguyen et al., 2023),

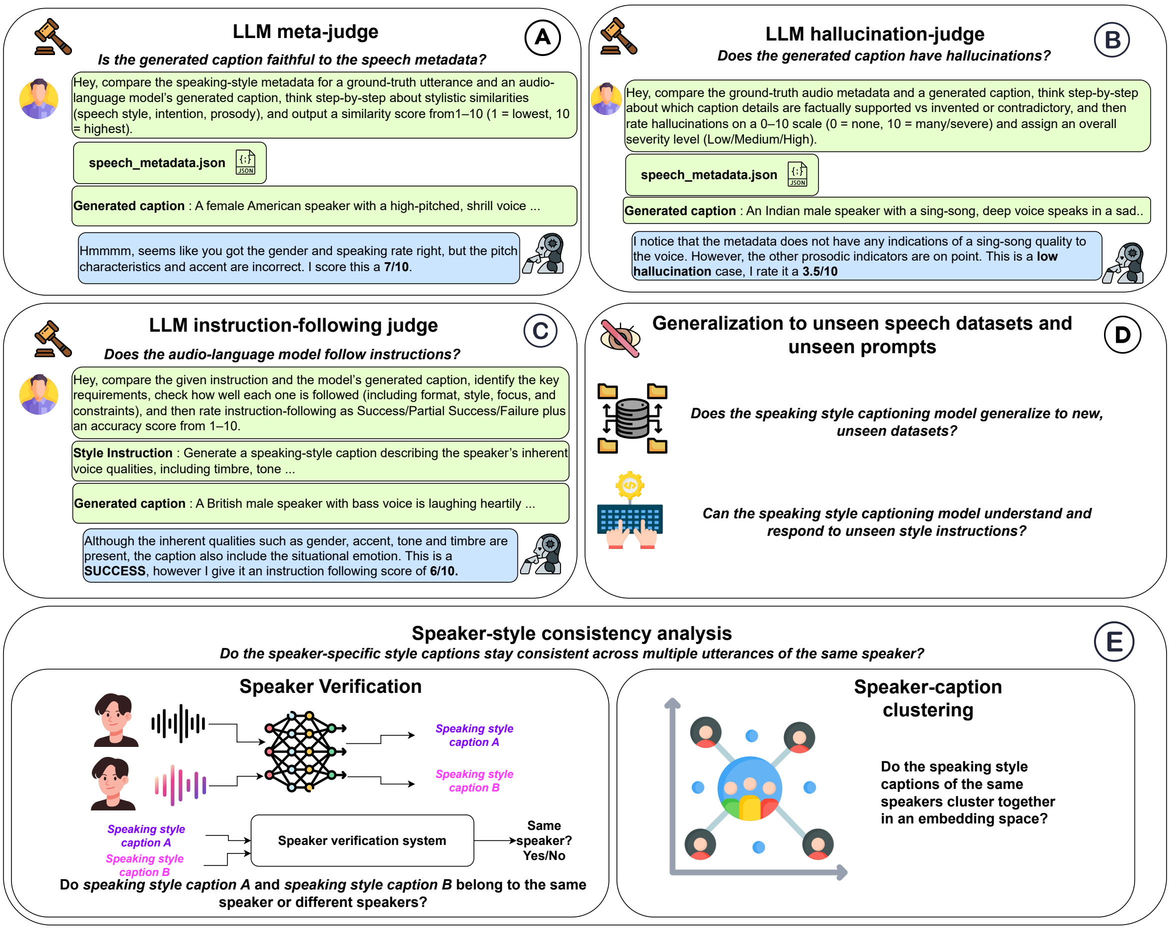


Figure 2: Overview of *StyleInstructCaps*, which evaluates models on metadata groundedness, instruction adherence, hallucination, speaker-style consistency, and generalization to unseen prompts and speech data.

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. Together, these datasets provide broad coverage of intrinsic and situational speaking-style attributes across diverse speakers and recording conditions.

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 B , 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 (Dubey et al., 2024) for the evaluation split, ensuring that benchmark evaluation does not benefit from exposure to captions generated by the same language model. Appendix B Figure 5 presents the captioning prompt, including explicit definitions of the six speaking-style dimensions, while Appendix B Table 6 reports instruction-caption pairs for the same utterance shown in Appendix B Figure 4.

3.3 Evaluation Framework

For the first four evaluation tasks, we adopt the LLM-as-a-judge paradigm, which has been widely used for evaluating open-ended text generation tasks (Zheng et al., 2023). Prior work has shown that LLM-based judgments often align better with human preferences than traditional automatic metrics such as ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015), BERTScore (Zhang et al., 2019), and n-gram based measurements (Mondshine et al., 2025).

3.3.1 Metadata Groundedness

In the first evaluation task (Figure 2, Part A), we assess the faithfulness of generated speaking-style captions by grounding them in the available speech metadata. We define **metadata groundedness** as the extent to which a generated caption is supported by and consistent with the metadata associated with the speech utterance. For this evaluation, we sample utterances from the *StyleInstructCaps-DB* evaluation split with an emphasis on metadata richness. We prioritize utterances that contain both intrinsic and situational annotations, first selecting those with non-empty situational tags and then completing the set from the remaining pool. To promote diversity and avoid over-representation, we cap the number of utterances per speaker at 10 and restrict any single accent to at most 25% of the evaluation set, resulting in 500 utterances. We refer to this subset as *StyleInstructCaps-Eval-500*. The distribution plots of accent, emotion, signal-to-noise ratio (SNR) and speech duration can be found in Appendix B Figure 6.

This task assesses an SSC models ability to perceive and represent fine-grained paralinguistic attributes from speech. Using the overall style summary instruction, an LLM judge compares the generated caption against PSC ground-truth metadata, explicitly evaluating speaker gender, accent, pitch, speaking rate, environment or noise level, and emotion or expression. Particular emphasis is placed on prosodic alignment across these attributes. The judge assigns a scalar similarity score on a 110 scale with a brief justification, and we report the mean score across utterances as the meta-score. The full evaluation prompt is provided in Appendix C, Figure 8.

3.3.2 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-Eval-500*.

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 complete prompt is provided in Appendix C Figure 9.

3.3.3 Instruction-Following

We evaluate the controllability of SSC models through an explicit instruction-following task (See Figure 2 Part C). Using the *StyleInstructCaps-Eval-500* 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. The mean value of this score is used to report the instruction-following ability score. The complete evaluation prompts are provided in Appendix C Figure 10 and 11.

3.3.4 Generalization to Unseen Instructions

Beyond the six speaking-style dimensions evaluated in *StyleInstructCaps-Eval-500*, we assess generalization to unseen instructions that differ from those used during training. We design 36 novel prompts that impose diverse constraints and output requirements, including restrictions on demographic inference, structured output formats such as JSON or tables, role-based perspectives, and explicit uncertainty handling. These prompts are intended to test whether instruction tuning leads to genuine instruction generalization rather than overfitting to seen patterns.

Evaluation is conducted on the VCTK corpus (Yamagishi et al., 2019), which is unseen during training. We select 16 speakers spanning diverse accents, with two utterances per speaker. Each utterance is paired with all 36 unseen instructions (Appendix C Figure 12), resulting in $16 \times 2 \times 36$ instruction-utterance pairs. This setup controls for speaker and accent variability while directly probing instruction-following generalization.

3.3.5 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 3.3.1 and 3.3.2, we assess metadata groundedness and hallucination with respect to the GigaSpeech ground-truth captions. We note that Speechcraft provides more limited metadata coverage than PSC.

3.3.6 Speaker-Caption 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). Accordingly, SSC models should produce captions that are consistent across a speakers utterances while remaining invariant to phonetic content and recording conditions.

To evaluate speaker-level consistency, we design two complementary (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 B 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 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., 2022), followed by attention-masked average pooling and L2 normalization. Speaker consistency is quantified using cosine similarity between caption embeddings, and performance is evaluated using Equal Error Rate (EER) and Area Under the ROC Curve (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 Experiment Setup

We use the open-source Qwen-3-32B model (Yang et al., 2025a) in thinking mode as the LLM judge, motivated by recent evidence that Qwen models offer strong evaluative reasoning and preference-judging capabilities (Li et al., 2025), while enabling reproducible, open evaluation compared to closed-source alternatives.

We evaluate six open-source audio-language models: Audio-Flamingo-3, Voxtral-Mini-3B, Voxtral-Small-24B, Qwen2-Audio-7B-Instruct, MERaLiON, and SALMONN (specifications in Appendix E, Table 8). 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 a multi-caption set with multiple instruction–caption pairs per utterance in a multi-task setup, SALMONN (*SIC-DB*)-MT. Additional training and inference details are in Appendix E. Data pre-processing follows the ParaSpeechCaps setup (Diwan et al., 2025b), and split statistics are in Appendix B, Table 7.

Table 1: Metadata groundedness, hallucination severity, and instruction-following performance evaluated on *StyleInstructCaps-Eval-500*. Blue shading indicates relative performance for scalar metrics.

Model name	LLM metadata groundedness		LLM hallucination severity (%)			Instruction following	
	Meta score↑	Hall.↓	Low	Mid	High	IFA score↑	IFA rate(%)↑
Oracle Caption (PSC)	9.93	0.069	99.5	0.5	0	-	-
Oracle caption (*SIC-DB*)	9.93	0.036	99.8	0.2	0	8.48	71.3
Audio-Flamingo-3	5.69	6.07	15.4	20.7	63.8	8.21	62.4
Voxtral-Mini-3B	7.56	3.12	48.6	35.0	16.4	8.15	68.0
Voxtral-Small-24B	7.20	1.58	73.5	16.8	9.68	7.51	67.8
Qwen2-Audio-7B-Instruct	5.73	5.58	19.1	27.9	53.0	6.58	53.6
MERaLiON	6.57	1.25	81.8	10.6	7.60	6.11	27.7
SALMONN	7.85	3.63	42.1	28.3	29.5	7.50	58.5
SALMONN (PSC)	9.24	4.77	21.2	46.3	32.5	7.98	52.9
SALMONN (*SIC-DB*) - ST	9.32	3.88	35.2	37.3	27.4	8.37	65.6
SALMONN (*SIC-DB*) - MT	9.32	3.43	41.3	38.2	20.5	8.62	75.0

Table 2: Speaker-style consistency evaluation using clustering and verification metrics on the ID-balanced set and the UD-VCTK set. Darker shading indicates stronger performance within each column. Higher is better for ARI, NMI, Purity, InterSep, and ROC–AUC, while 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	0.0092	0.0461	0.1	0.0463	0.476	0.532	0.0113	0.0479	0.121	0.0574	0.463	0.556
Voxtral-Mini-3B	0.0326	0.122	0.152	0.133	0.399	0.643	0.0034	0.0282	0.103	0.0631	0.486	0.519
Voxtral-Small-24B	0.0666	0.182	0.186	0.157	0.363	0.674	0.0108	0.052	0.119	0.0758	0.48	0.532
Qwen2-Audio-7B-Instruct	-0.0018	0.0101	0.0715	0.0258	0.498	0.499	-0.0023	0.0069	0.0825	0.025	0.496	0.498
MERaLiON	0.0814	0.276	0.213	0.233	0.428	0.604	0.0393	0.146	0.167	0.169	0.476	0.534
SALMONN	0.351	0.598	0.511	0.29	0.225	0.854	0.145	0.324	0.303	0.205	0.32	0.749
SALMONN (PSC)	0.297	0.541	0.491	0.271	0.231	0.861	0.157	0.34	0.312	0.209	0.32	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	0.32	0.749
SALMONN(*SIC-DB*) - MT	0.357	0.59	0.517	0.279	0.212	0.876	0.156	0.341	0.298	0.222	0.282	0.795

5 Results and Discussion

Table 1 reports the results for the first three evaluation tasks – metadata groundedness, hallucination severity and instruction-following, all evaluated by the LLM judge. 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. This indicates that task-specific training is necessary for improvements on the SSC benchmark. Appendix D Figure 13 provides an example of judgment for the metadata groundedness evaluation.

We observe that **models trained with relevant audio encoder and tasks to SSC perform better than models that have only seen ASR in training** - in which case the speech understanding is limited to the *what* is being said and cannot generalize to *how* it is said. Although Audio-Flamingo-3 achieves state-of-the-art performance on general audio benchmarks such as MMAU (Sakshi et al., 2024) and MMAR (Ma et al., 2025), its low LLM metascore and elevated hallucination indicate limited effectiveness on SSC. In contrast, SALMONN-despite a smaller LLM backbone and reduced training scale-leverages a more suitable audio encoder combining BEATS (Chen et al., 2022b) and Whisper (Radford et al., 2023), along with SSC-relevant training tasks (SQA, GR, SER), enabling it to consistently outperform other audio-language models in our evaluation. Finally, models that are explicitly trained for SSC show increased understanding of paralinguistics as they achieve high metascores and lower hallucination scores.

We next examine instruction-following performance. As shown in Table 1, instruction-following shows little correlation with metadata groundedness or hallucination severity, demonstrating that these metrics capture distinct model competencies. Performance is primarily driven by the diversity of tasks seen during instruction tuning, the capacity of the underlying LLM backbone, and hyper-parameter choices that prevent overfitting to narrow instruction distributions. Notably, models such as Qwen2-Audio and the Voxtral series achieve strong instruction-following scores despite weak faithfulness-oriented performance. **This dissociation indicates that generic instruction tuning**

Table 3: Evaluation on unseen datasets and prompting conditions.

Model	Unseen Eval-set		OOD Prompts	
	Meta score↑	Hall.↓	IF score↑	IF (%)↑
Audio-Flamingo-3	4.26	5.05	6.12	0.323
Voxtral-Mini-3B	5.21	5.93	6.51	0.395
Voxtral-Small-24B	3.87	5.65	5.85	0.437
Qwen2-Audio-7B-Instruct	5.35	5.78	6.76	0.434
MERaLiON	4.45	5.18	5.38	0.260
SALMONN	8.26	4.91	6.67	0.357
SALMONN (PSC)	8.69	4.63	6.73	0.332
SALMONN (*SIC-DB*) - ST	8.32	4.49	7.03	0.391
SALMONN (*SIC-DB*) - MT	8.69	4.53	7.02	0.389

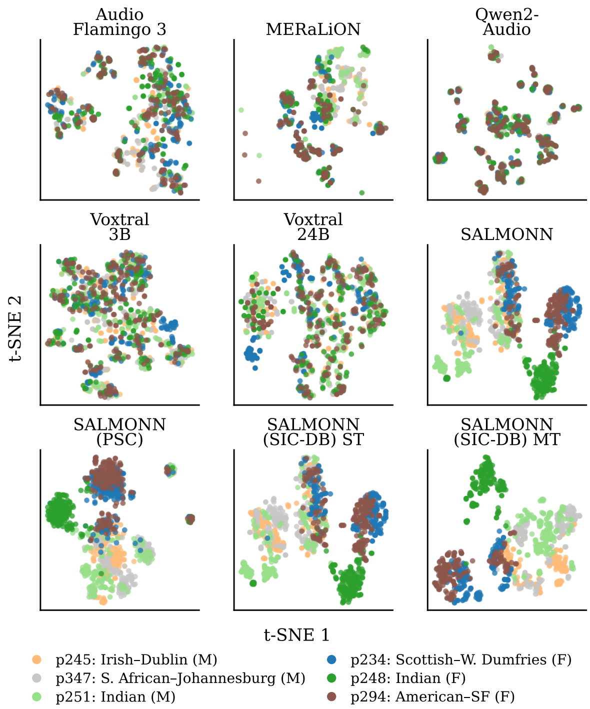


Figure 3: t-SNE visualization of speaker-style caption embeddings on UD-VCTK (6 speaker subset)

suffices to produce stylistically well-formed captions but fails to ground them in speech-specific paralinguistic attributes. Consequently, while SSC-specific training is not required for instruction compliance, it is essential for accurately describing how speech is delivered. This trend persists under evaluation on unseen data and prompts (Table 3). While task-specific SSC models substantially outperform general-purpose audio-language models in LLM-judge meta groundedness and hallucination metrics, instruction-following performance remains largely unaffected by SSC-specific training.

Table 2 reveals clear and consistent trends in speaker-style consistency across both the ID-balanced set and UD-VCTK sets. **SSC-specialized training yields substantial gains over general-purpose audio–language models**, with all SALMONN-based variants outperforming Audio-Flamingo-3, Voxtral, Qwen2-Audio-7B-Instruct, and MERaLiON across clustering (ARI, NMI, Purity, InterSep) and verification (EER, AUC) metrics. These gains indicate that SSC training enables captions to 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) clearly demonstrate the advantage of task-specific training over general-purpose models, as evidenced by more coherent and well-separated speaker clusters. A comprehensive t-SNE analysis covering all speakers across all models is provided in Appendix F Figure 14.

Compared to SALMONN (PSC), explicit style supervision consistently improves speaker consistency, demonstrating the benefit of rich speaking-style annotations even without instruction-based control. 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, the results demonstrate that SSC-specific and instruction-diverse training is essential for learning speaker-consistent speaking-style representations that generalize across domains.**

6 Conclusions

We introduced *StyleInstructCaps*, a benchmark for controllable SSC that formulates as an instruction-following audio-language task and evaluates models across five complementary axes. Through experiments on six state-of-the-art audio-language models, we showed that task-aligned and instruction-diverse training is critical for SSC, substantially improving metadata groundedness, reducing hallucination, strengthening speaker-style consistency, and enabling robust generalization to unseen speakers, datasets, and prompting conditions beyond generic large-scale audio-language pretraining. With this work, we set a standardized evaluation foundation for SSC, transforming it from a loosely defined capability into a measurable research task.

Limitations

This study is subject to several practical constraints inherent to speaking-style captioning. First, large-scale speech corpora with rich, fine-grained, per-utterance speaking-style metadata remain limited in availability, which naturally bounds the diversity and scale of supervised SSC training and evaluation. Second, evaluating open-ended style descriptions necessarily relies on LLM-based judgment, and while recent LLM judges offer strong reasoning capabilities, their assessments are influenced by both the judge model itself and the coverage of the underlying datasets. Finally, despite efforts to reduce prompt sensitivity through neutral prompt generation and evaluation over multiple prompts, performance may still vary with alternative prompt formulations or decoding strategies.

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. 2022a. 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, and Furu Wei. 2022b. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.

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. *arXiv 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. *arXiv 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. 2025a. Scaling rich style-prompted text-to-speech datasets. *Preprint*, arXiv:2503.04713.

Anuj Diwan, Zhisheng Zheng, David Harwath, and Eunsol Choi. 2025b. Scaling rich style-prompted text-to-speech datasets. *Preprint*, arXiv:2503.04713.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

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.

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.

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.

Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. 2025. Evaluating scoring bias in llm-as-a-judge. *arXiv preprint arXiv:2506.22316*.

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.

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, and 1 others. 2025. Desta2. 5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *arXiv preprint arXiv:2507.02768*.

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, and 1 others. 2025. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*.

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. 2024. 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.

Jing Peng, Yucheng Wang, Yangui Fang, Yu Xi, Xu Li, Xizhuo Zhang, and Kai Yu. 2024. A survey on speech large language models. *arXiv preprint arXiv:2410.18908*.

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. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

Christoph Schuhmann, Robert Kaczmarczyk, Gollam Rabby, Felix Friedrich, Maurice Kraus, Kourosh Nadi, Huu Nguyen, Kristian Kersting, and Sören Auer. 2025. Emonet-voice: A fine-grained, expert-verified benchmark for speech emotion detection. *arXiv 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. *arXiv preprint arXiv:2407.02005*.

Bo-Hao Su, Hui-Ying Shih, Jinchuan Tian, Jiatong Shi, Chi-Chun Lee, Carlos Busso, and Shinji Watanabe. 2025. Reasoning beyond majority vote: An explainable speechlm framework for speech emotion recognition. *arXiv preprint arXiv:2509.24187*.

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.

Andreas Triantafyllopoulos, Anton Batliner, and Björn W Schuller. 2025. Charting 15 years of progress in deep learning for speech emotion recognition: A replication study. *arXiv preprint arXiv:2508.02448*.

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. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv 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. *arXiv preprint arXiv:2509.16589*.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, and 1 others. 2025. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

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. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shu-wen Yang, Ming Tu, Andy T Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. 2025b. Paras2s: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. *arXiv preprint arXiv:2511.08723*.

Jian Zhang, Linhao Zhang, Bokai Lei, Chuhan Wu, Wei Jia, and Xiao Zhou. 2025a. Wildspeech-bench: Benchmarking audio llms in natural speech conversation. *arXiv preprint arXiv:2506.21875*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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*.

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 75137522, 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., 2025a)	2.9k

Table 4: Summary of existing datasets relevant to SSC.

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 interpretation.s	Task accuracy, BLEU (for generation), MOS-style human ratings.
ParaS2SBench(Yang et al., 2025b)	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., 2025)	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 (Wu et al., 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/Rš/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 5: Benchmarks for Speech Paralinguistics Understanding

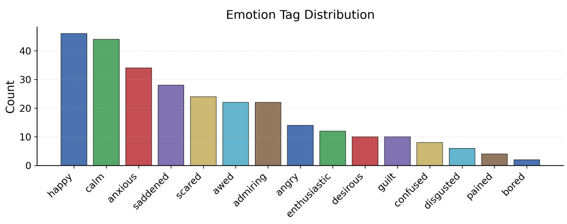
B Dataset details


```
{
"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": " o, li? ju w so d æt ðæt. a lv tu dst st hi nd lsn tu ju ple. t’s so læks. a dst kæn’ti
bliv ha tælntd ju .",
"audio_path": "datasets/ears/p047/emo_amazement_freeform.wav",
}
```

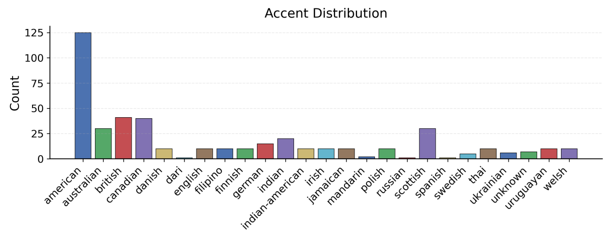

Figure 4: Example metadata object from EARS used in our captioning/evaluation pipeline.

```
[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>"
}
```

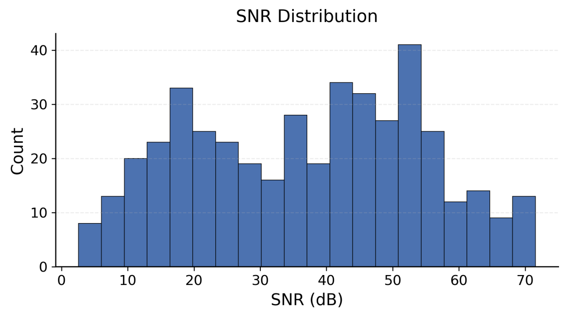
Figure 5: Prompt template for generating six style-specific captions from an input metadata object.



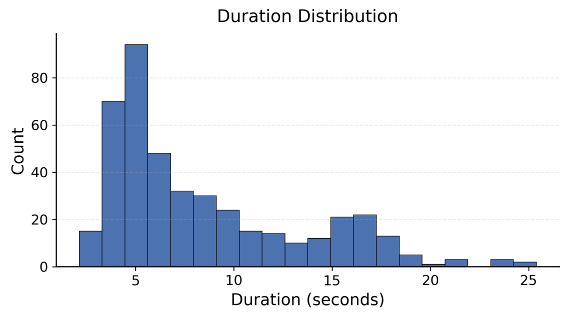
(a) Emotion



(b) Accent

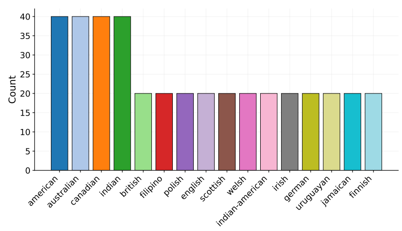


(c) SNR

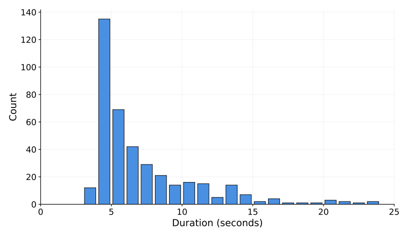


(d) Duration

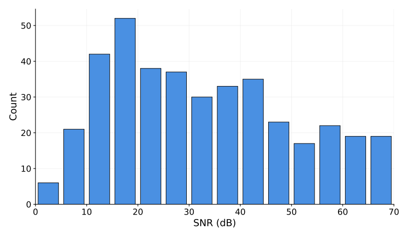
Figure 6: Distributions of emotion (a), accent (b) signal-to-noise (SNR) ratio (c), and speech duration (d) in *StyleInstructCaps*-Eval-500.



(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 6: Style-specific instructions and example generated captions.

Table 7: 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-Eval-500	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

C 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 8: 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 9: 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 10: 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 11: Instruction-following evaluation prompt (Part 2): scoring rubric, edge cases, and required output format.

```
{
"unseen_prompt_1": "Establish the speakers 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 speakers 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 speakers 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, äroural̈: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 speakers 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 directors 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 12: Unseen instruction prompts used to probe diverse instruction-following behaviors for speaking-style captioning.

D 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 13: Example showing an LLM judgment (top) and the corresponding ground-truth metadata vs. generated caption (bottom).

E 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/
SALMONN-7B	Evaluated	https://github.com/bytedance/SALMONN/tree/salmonn
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
Qwen-3-32B	LLM-Judge	https://huggingface.co/Qwen/Qwen3-32B

Table 8: Specification and sources of the employed models.

Training and inference details. We fine-tune all audio–language backbones 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`.

F Speaker Caption clustering plots

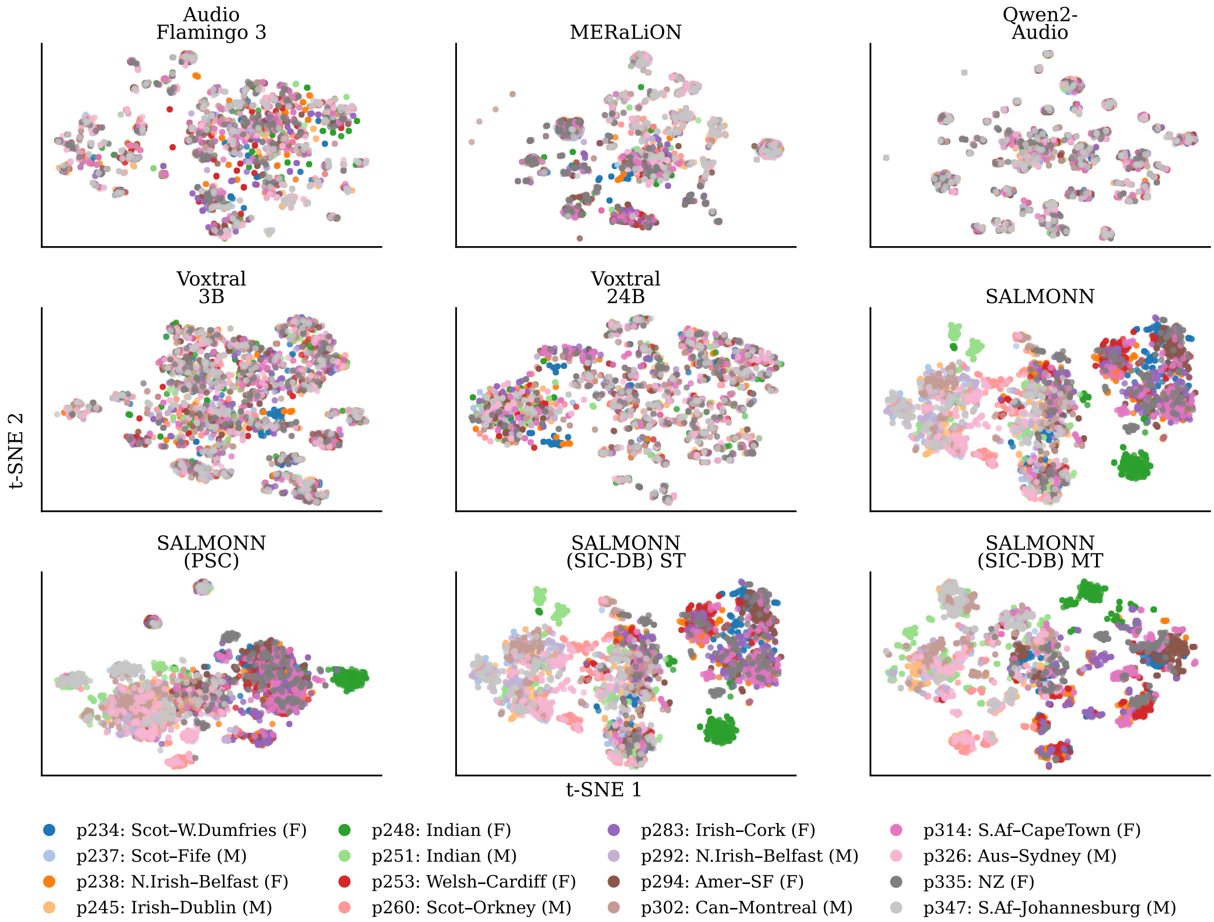


Figure 14: t-SNE visualization of speaker-style caption embeddings on UD-VCTK (all speakers)

G 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.

H 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.