Towards Fine-Grained and Multi-Granular Contrastive Language-Speech Pre-training

Anonymous ACL submission

Abstract

Modeling fine-grained speaking styles remains challenging for language-speech representation pre-training, as existing speech-text models are typically trained with coarse captions or task-specific supervision, and scalable fine-grained style annotations are unavailable. We present FCaps, a large-scale dataset with fine-grained free-text style descriptions, encompassing 47k hours of speech and 19M fine-grained captions annotated via a novel end-to-end pipeline that directly grounds detailed captions in audio, thereby avoiding the error propagation caused by LLM-based rewriting in existing cascaded pipelines. Evaluations using LLM-as-a-judge demonstrate that our annotations surpass existing cascaded annotations in terms of correctness, coverage, and naturalness. Building on FCaps, we propose CLSP, a contrastive language-speech pre-trained model that integrates global and fine-grained supervision, enabling unified representations across multiple granularities. Extensive experiments demonstrate that CLSP learns fine-grained and multi-granular speech-text representations that perform reliably across global and fine-grained speech-text retrieval, zero-shot paralinguistic classification, and speech style similarity scoring, with strong alignment to human judgments. All resources will be made publicly available.

1 Introduction

Speaking style conveys rich paralinguistic information beyond lexical content, encompassing both speaker-intrinsic characteristics (e.g., gender, age, and accent) and temporally varying traits (e.g., intonation, emotion, and expressiveness). In this sense, speaking style is inherently multi-scale: it can be summarized at the global, utterance level without explicit temporal structure, or described through fine-grained stylistic variation within an utterance.

However, modeling speaking style remains challenging. Existing approaches (Ma et al., 2024) typically rely on utterance-level, discrete labels, which limit expressive diversity and fail to capture the temporal structure of speech. Evaluating speaking style is likewise challenging. Widely used subjective human judgments (Yang et al., 2025a) suffer from limited inter-rater consistency and are hard to scale. Recent automated alternatives based on large audio language models (OpenAI, 2024; Comanici et al., 2025) incur huge costs. As a result, there remains substantial scope for fine-grained modeling and scalable evaluation of speaking style in speech–text representation learning.

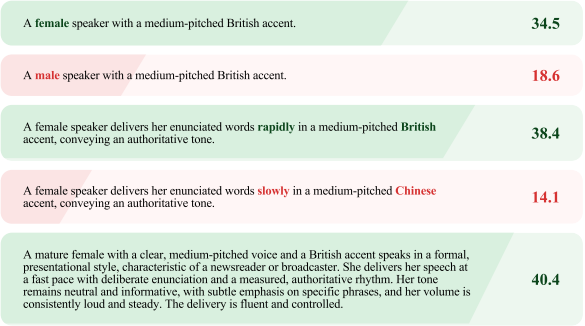

Figure 1: Multi-granular speech style caption similarity scoring for the same speech input by CLSP. Positive captions (green) receive higher scores, while hard negatives (red), despite mainly textual overlap, receive markedly lower scores due to attribute mismatches.

A central bottleneck is the lack of scalable, reliable, and fine-grained style annotations. Existing speech style-captioned datasets (Jin et al., 2024; Diwan et al., 2025; Wang et al., 2025) predominantly use cascaded annotation pipelines, in which speech is first labeled with discrete attributes and then rewritten into free-text descriptions by large language models, often introducing error propagation and semantic misalignment. More fundamentally, these intermediate discrete attributes impose a severe information bottleneck, compressing rich, continuous, and temporally varying paralinguistic information into a finite set of predefined categorical tags and leading to substantial information loss.

Table 1: Comparison of *open-source* English speech style-captioned datasets. Following (Wang et al., 2025), I1-I5 denote intrinsic speaker traits: age (I1), gender (I2), timbre/texture (I3), mean pitch (I4), and accent (I5). S1-S4 represent situational traits: speaking rate (S1), emotion (S2), expressivity (S3), and volume (S4).

Dataset	I1	I2	I3	I4	I5	S1	S2	S3	S4	Free-Text	End-to-End	# Hours
Expresso (Nguyen et al., 2023)	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	47
EARS (Richter et al., 2024)	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗	✗	60
PromptSpeech (Guo et al., 2023)	✗	✓	✗	✓	✗	✓	✓	✗	✓	✗	✗	0.3k
TextrolSpeech (Ji et al., 2024)	✗	✓	✗	✓	✗	✓	✓	✗	✓	✗	✗	0.3k
VccmDataset (Ji et al., 2025)	✗	✓	✗	✓	✗	✓	✓	✗	✓	✗	✗	0.3k
DreamVoiceDB (Hai et al., 2024)	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	0.3k
LibriTTS-P (Kawamura et al., 2024)	✓	✓	✓	✓	✗	✓	✗	✓	✓	✗	✗	0.6k
SpeechCraft (Jin et al., 2024)	✓	✓	✗	✓	✗	✓	✓	✗	✓	✓	✗	2.4k
ParaSpeechCaps (Diwan et al., 2025)	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	2.9k
ParlerTTS (Lyth and King, 2024)	✗	✓	✗	✓	✓	✓	✗	✓	✗	✓	✗	44.5k
CapSpeech (Wang et al., 2025)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	33.6k
FCaps	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	47.1k

With these perspectives in mind, we introduce FCaps, a large-scale speech dataset paired with fine-grained free-text speaking-style descriptions, constructed via an end-to-end annotation pipeline that directly grounds style captions in audio and incorporates agentic verification to improve annotation quality. We show that our annotations achieve substantially higher quality than cascaded annotations in terms of correctness, coverage, and naturalness, as assessed by LLM-as-a-judge. Building on FCaps, we further propose CLSP, a contrastive language–speech pre-trained model that leverages fine-grained and multi-granular supervision to learn unified speech–text representations across different levels of granularity (illustrated in Figure 1). Extensive experiments demonstrate strong performance across multiple tasks, including global and fine-grained speech–text retrieval, zero-shot paralinguistic classification, and speech style similarity scoring aligned with human judgments.

Our contributions are fourfold:

- We present FCaps, the largest dataset to date for fine-grained free-text speaking-style descriptions, with 47k hours of speech and 19M captions.
- We propose an end-to-end pipeline for fine-grained style annotation, substantially improving annotation quality over existing approaches.
- We present CLSP, a speech–text dual-encoder trained with fine-grained and multi-granular contrastive supervision, enabling unified representation learning across multiple granularities.
- We demonstrate that CLSP learns robust and generalizable speech–language representations that perform reliably across multiple tasks, with strong alignment to human judgments.

2 Related Work

2.1 Speech Style-Captioned Datasets

Table 1 summarizes key differences between our dataset and prior work. Existing speech style-captioned datasets (Guo et al., 2023; Ji et al., 2024; Lyth and King, 2024; Jin et al., 2024; Diwan et al., 2025; Wang et al., 2025) typically adopt cascaded annotation pipelines, where speech is labeled with discrete tags and then rewritten into natural-language captions by LLMs. This cascaded design introduces an information bottleneck by compressing rich, continuous paralinguistic cues into a finite set of discrete tags, resulting in coarse captions that provide limited support for fine-grained modeling.

2.2 Contrastive Language–Speech Pre-training

Contrastive learning has emerged as a powerful paradigm for multimodal pre-training, demonstrating strong performance across image–text (Radford et al., 2021) and audio–text (Elizalde et al., 2023; Wu et al., 2023). In the speech domain, existing speech–text contrastive models are predominantly trained with coarse-grained captions or task-specific supervision, offering limited modeling of the temporal and narrative structure of speech. GLAP (Dinkel et al., 2025) learns general representations across speech, audio, and music, but relies on paired transcriptions for speech, providing primarily lexical-level supervision. RA-CLAP (Sun et al., 2025) adopts coarse-grained captions. ParaCLAP (Jing et al., 2024) focuses on emotion-centric supervision. Overall, modeling fine-grained and multi-granular speaking styles in a unified manner remains underexplored.

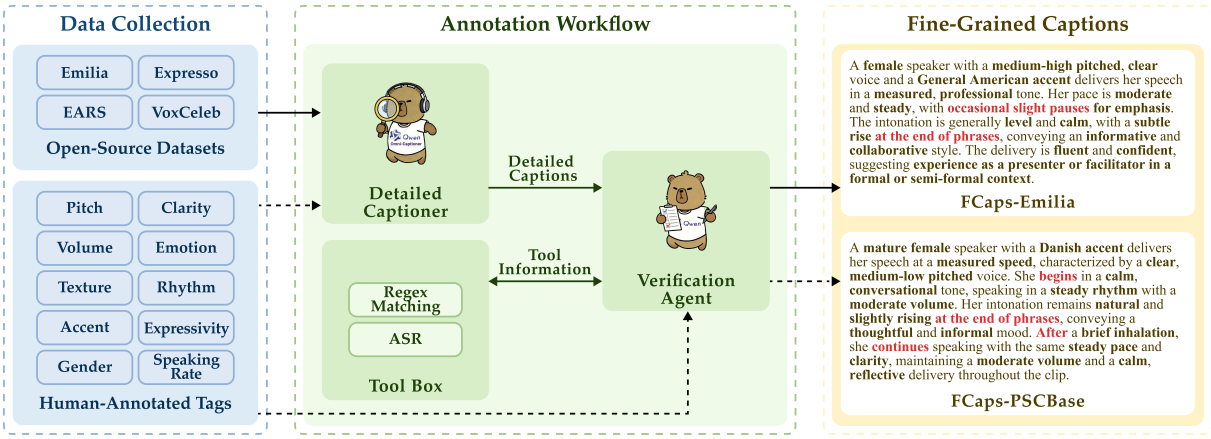

Figure 2: Overview of our end-to-end annotation pipeline for generating fine-grained captions, consisting of a detailed captioner and agentic verification with specialist tools. Solid lines indicate the construction process for FCaps-Emilia, and dashed lines indicate additional processes for FCaps-PSCBase. In the example fine-grained captions, speaker-related traits are highlighted in bold and narrative structure in red.

3 FCaps Dataset Construction

3.1 Caption Taxonomy: Global and Fine-Grained

We define two types of textual supervision:

- **Global captions** provide a holistic description of the speech that summarizes speaker-related attributes, encompassing intrinsic traits tied to a speaker's identity and stable across utterances, and situational traits that may vary across utterances. Such descriptions are atemporal in nature and do not narrate intra-utterance variations.
- **Fine-grained captions** extend beyond a holistic speaker profile by providing a temporal and narrative structure that tracks within-clip dynamics such as style shifts, prosodic variations, emphasis patterns, and non-verbal vocalizations, and may further encode the speaker's delivery style, communicative role, and communicative intent.

Together, they provide multi-granular views of the same speech signal, thereby supporting fine-grained contrastive learning of a unified representation across multiple granularities.

3.2 End-to-End Annotation Pipeline

Figure 2 illustrates our proposed end-to-end annotation pipeline for generating fine-grained speech captions. Given a speech clip as input, the pipeline consists of two processes: detailed caption generation and agentic verification. A multimodal captioner generates multiple candidate captions for the input speech, optionally taking available human-annotated tags. A verification agent then evaluates each candidate using a predefined checklist and a toolbox of specialist tools, and decides whether the candidate is retained or filtered.

Detailed Captioning Detailed captioning, also referred to as detailed perception, aims to generate fine-grained descriptions of an audio or video segment, emphasizing maximal perceptual detail within the clip duration. In this work, we employ Qwen3-Omni-30B-A3B-Captioner[1] as the detailed captioner, which produces detailed and low-hallucination captions (Ma et al., 2025), capturing speaker emotions, layered intentions, cultural context, implicit cues, and additionally supports non-speech sound recognition and analysis.

Although the captioner is designed to operate without prompting, its default outputs often include spoken-content transcription, environmental sound descriptions, and audio quality assessments, which are irrelevant for speaker-centric contrastive learning. We therefore apply user prompt conditioning to suppress such content. Leveraging the strong instruction-following capability inherited from the base model, the captioner can reliably adhere to these constraints, thereby steering caption generation towards relatively concise and speaker-focused descriptions. A qualitative case study on how different prompt compositions influence captioner outputs is provided in Appendix B.

Multi-Positive Captioning To construct multiple positive textual views of the same speech clip, we perform caption generation multiple times using different random seeds. This yields a set of

[1] https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Captioner

captions that are all grounded in the same speech signal, while differing in lexical choice, descriptive focus, and narrative structure. Previous work (Jin et al., 2024; Wang et al., 2025) adopts text-based data augmentation that rewrites captions purely in the textual modality using LLMs with complex instructions. Such approaches suffer from instruction non-compliance and introduce distorted attributes or hallucinations in a non-trivial fraction of cases, as the rewriting process is not conditioned on the original speech signal. By contrast, our approach conditions the generation process directly on the audio input, producing multiple acoustically consistent yet semantically non-identical positive views that are well suited for contrastive learning. A case study is provided in Appendix C.

Agentic Verification To improve the reliability of generated captions, we introduce an agentic verification process centered on a text-based reasoning model, Qwen3-30B-A3B-Thinking-2507[2].

The verification agent evaluates each candidate caption according to a predefined checklist and discards it if *any* item in the checklist is violated. The checklist targets common failure modes in detailed captioning by examining whether a caption: (1) includes descriptions of background sounds, environmental noise, or audio quality; (2) contains explicit declarations about the absence of certain elements; (3) contains spoken-content transcription without attached style descriptions; or (4) fails to appropriately incorporate human-annotated tags when available. In addition, for speech clips from EARS (Richter et al., 2024), Expresso (Nguyen et al., 2023), and VoxCeleb (Nagrani et al., 2017), the agent enforces a clip-level constraint requiring a single speaker with a single role. If a caption indicates multiple speakers or a single speaker assuming multiple roles, the corresponding speech clip and all associated captions are discarded.

To support these judgments, the agent leverages specialist tools, including (1) rule-based pattern matching implemented via Python regular expressions, (2) access to the speech transcription, and (3) access to human-annotated tags when available. Based on the aggregated evidence, the agent makes a binary retain-or-filter decision. This agentic verification process enables systematic quality control, effectively eliminating captions containing extraneous content or incomplete grounding.

[2] https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507

3.3 FCaps-Emilia

FCaps-Emilia is constructed from the Emilia corpus (He et al., 2024). For each speech segment, the detailed captioner is run five times to generate candidate fine-grained captions. Given the large scale of the audio sources, the verification agent retains a single verified fine-grained caption per utterance. In total, FCaps-Emilia comprises 18,131,371 fine-grained captions, covering 46,787 hours of speech. FCaps-Emilia does not include global captions.

3.4 FCaps-PSCBase

FCaps-PSCBase is built upon the PSC-Base corpus (Diwan et al., 2025), incorporating audio clips from EARS, Expresso, and VoxCeleb, along with additional human-annotated tags and captions. We adopt the captions provided by PSC-Base as global captions and apply rule-based normalization to mitigate common artifacts arising from LLM rewriting and fuzzy matching to correct spelling errors. We incorporate speaking rate, accent, and situational tags to guide fine-grained caption generation. For each utterance, the detailed captioner is run 20 times to obtain candidate fine-grained captions. After agentic verification, between 5 and 14 verified captions per utterance are retained as multi-positive views. Overall, FCaps-PSCBase comprises 140,602 global captions and 930,917 fine-grained captions, spanning 267 hours of speech.

4 CLSP Model

4.1 Model Architecture

As shown in Figure 3, CLSP adopts the dual-encoder architecture of CLAP, where speech and text are processed by separate encoders, followed by MLP projection to map two modalities into a shared embedding space.

Speech and Audio Unified Encoder SPEAR-XLarge[3] is used as the speech and audio unified encoder, with representations extracted from the final encoder layer. SPEAR (Yang et al., 2025b) is a unified self-supervised representation model for both speech and general audio, and achieves state-of-the-art performance across a range of speech and audio benchmarks, making it well-suited for capturing fine-grained acoustic and paralinguistic cues in speaker-centric contrastive learning.

[3] https://huggingface.co/marcoyang/spear-xlarge-speech-audio

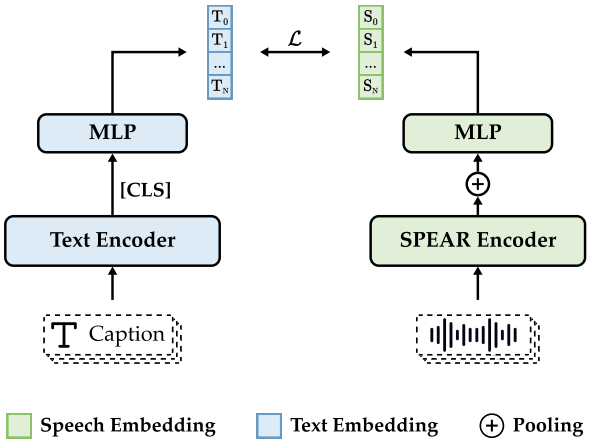

Figure 3: Overview of CLSP.

Text Encoder RoBERTa-base[4] is used as the text encoder. We adopt variable-length inputs with a maximum of 512 tokens to accommodate both global captions and fine-grained captions of varying granularity. The sentence-level representation is obtained from the final-layer `[CLS]` token.

4.2 Fine-Grained and Multi-Granular Contrastive Language–Speech Pre-training

We adopt a two-stage curriculum for speech–text representation learning with fine-grained and multi-granular contrastive supervision. The training focus progressively shifts from pure fine-grained alignment to a balance between cross-granularity generalization and robust fine-grained discrimination. In the first stage, speech and text are aligned using standard contrastive learning on large-scale data paired with fine-grained captions. In the second stage, we introduce multi-positive contrastive learning with diverse global and fine-grained captions at multiple granularities.

Stage One Given a speech clip $\mathbf{x}$ and its paired tokenized fine-grained caption $\mathbf{y}_{\mathrm{F}}$, the speech encoder produces frame-level representations that are aggregated via mean pooling over time, followed by MLP projection and ℓ_2 normalization to obtain a speech embedding $\mathbf{s} \in \mathbb{R}^d$. For text, we take the final-layer `[CLS]` hidden state from the text encoder, followed by an MLP projection and ℓ_2 normalization to obtain a text embedding $\mathbf{t}_{\mathrm{F}} \in \mathbb{R}^d$. Here, d is the embedding space dimensionality. We use a symmetric InfoNCE (He et al., 2020) loss, with each paired speech and text forming a positive example, and all other non-matching pairs within the same batch serving as negatives:

$$\mathcal{L} = -\frac{1}{2N}\sum_{i=1}^{N}\Bigg(\log\frac{\exp(\mathbf{s}_i\cdot\mathbf{t}_{\mathrm{F}i}/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{s}_i\cdot\mathbf{t}_{\mathrm{F}j}/\tau)} + \log\frac{\exp(\mathbf{t}_{\mathrm{F}i}\cdot\mathbf{s}_i/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{t}_{\mathrm{F}i}\cdot\mathbf{s}_j/\tau)}\Bigg), \quad (1)$$

where N is the batch size and τ is a learnable temperature. Since all embeddings are ℓ_2-normalized, dot products are equivalent to cosine similarity.

Stage Two We adopt a symmetric multi-positive InfoNCE loss, implemented as cross-entropy with soft targets. Given a batch of N speech samples $\{\mathbf{x}_i\}_{i=1}^{N}$, each paired with two tokenized captions $\{\mathbf{y}_i, \hat{\mathbf{y}}_i\}$, we obtain a speech embedding $\mathbf{s}_i \in \mathbb{R}^d$ and two text embeddings $\mathbf{t}_i, \hat{\mathbf{t}}_i \in \mathbb{R}^d$ in the same manner as in Stage One. We stack the speech embeddings as $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_N] \in \mathbb{R}^{N\times d}$ and the text embeddings as $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_N, \hat{\mathbf{t}}_1, \ldots, \hat{\mathbf{t}}_N] \in \mathbb{R}^{2N\times d}$, and compute the similarity logits $\mathbf{L} = \mathbf{S}\mathbf{T}^\top \in \mathbb{R}^{N\times 2N}$. For audio-to-text direction, we define a soft target distribution $\mathbf{D} \in \mathbb{R}^{N\times 2N}$ that assigns probability mass λ and $1-\lambda$ to two paired texts, and zero to all others:

$$M_{i,j} = \begin{cases} \lambda, & \text{if } j = i, \\ 1-\lambda, & \text{if } j = i+N, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

We set $\lambda = 0.5$ based on the ablation study reported in Appendix G.2. For text-to-audio direction, each text embedding has a single speech, yielding target distribution $\mathbf{D}' \in \mathbb{R}^{2N\times N}$:

$$N_{j,i} = \begin{cases} 1, & \text{if } j = i \text{ or } j = i+N, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

The loss is defined as the average of two directions:

$$\mathcal{L} = \frac{1}{2}\Big(\mathrm{CE}(\mathbf{L}/\tau, \mathbf{D}) + \mathrm{CE}(\mathbf{L}^\top/\tau, \mathbf{D}')\Big), \quad (4)$$

where $\mathrm{CE}(\cdot,\cdot)$ denotes cross-entropy and τ is a learnable temperature parameter.

For each training step, the model samples one of two tasks according to a task scheduler:

- **Task 1**: each speech sample is paired with a global caption and a fine-grained caption, encouraging cross-granularity generalization.
- **Task 2**: each speech sample is paired with two distinct fine-grained captions, improving fine-grained discrimination via semantic consistency.

[4] https://huggingface.co/FacebookAI/roberta-base

We explore both static and dynamic task schedulers. At training step t, Task 1 is sampled with probability p_t, while Task 2 is sampled with probability $1 - p_t$. For the static scheduler, $p_t = p_0$ is fixed. For the dynamic scheduler, p_t decreases linearly from p_0 to $p_{\min}$ over T training steps and remains fixed thereafter:

$$p_t = \max\left(p_{\min},\ p_0 - \frac{t}{T}(p_0 - p_{\min})\right). \quad (5)$$

The best-performing strategy uses a dynamic scheduler with $p_0 = 0.95$, $p_{\min} = 0.50$, and $T = 10{,}000$. We provide ablation studies of different task schedulers in Appendix G.3.

5 Experiments

5.1 Annotation Quality of FCaps

Evaluation Setup We evaluate the annotation quality of FCaps by comparing captions generated by our end-to-end annotation pipeline with those produced by a representative cascaded annotation pipeline. Specifically, we randomly sample 1,000 audio clips from FCaps-Emilia along with one caption per audio generated by our end-to-end pipeline. For the same set of audio clips, we obtain the corresponding captions from PSC-Scaled (Diwan et al., 2025) as the cascaded baseline, which consists of automatically predicted discrete style labels, filtered by Gemini 1.5 Flash (Reid et al., 2024), and rewritten into natural-language style captions by Mistral-7B-Instruct-v0.2 (Jiang et al., 2023).

LLM-as-Judges We employ Gemini 3 Pro (gemini-3-pro-preview, November 2025) as a multimodal judge to evaluate caption quality. Each evaluation query consists of an audio clip and its two corresponding captions (cascaded *vs.* end-to-end), and assigns absolute scores to each caption along three dimensions: (1) audio-grounded *correctness*, measuring factual consistency with the audible content; (2) *coverage*, assessing whether speaking-style attributes are adequately captured; and (3) *naturalness*, evaluating fluency, grammaticality, and human-likeness. All evaluations follow an identical prompt and scoring rubric, detailed in Appendix D. Scores are averaged over 5 runs with randomized order to reduce variance.

Results Table 2 reports Gemini 3 Pro scores for cascaded and end-to-end caption annotations. Our end-to-end annotations consistently outperform cascaded annotations across all three dimensions by a large margin. Figure 4 further shows that end-to-end captions outperform cascaded captions in the majority of cases across all evaluation dimensions, with especially large gains in coverage and correctness. These gaps highlight the information bottleneck introduced by intermediate discrete attributes in cascaded pipelines, which irreversibly compress rich paralinguistic information. By contrast, our end-to-end pipeline yields better alignment with the audible content, more comprehensive coverage of speaking-style attributes, and more fluent, human-like descriptions.

Table 2: Comparison of end-to-end and cascaded caption annotations evaluated by Gemini 3 Pro.

Pipeline	Correctness	Coverage	Naturalness	Avg.
Cascaded	$3.30_{\pm 0.05}$	$3.10_{\pm 0.02}$	$4.15_{\pm 0.05}$	$3.51_{\pm 0.04}$
End-to-end	$\mathbf{4.42}_{\pm 0.04}$	$\mathbf{4.55}_{\pm 0.03}$	$\mathbf{4.92}_{\pm 0.02}$	$\mathbf{4.63}_{\pm 0.03}$

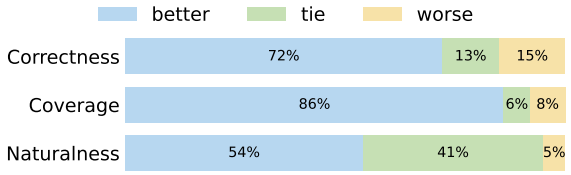

Figure 4: Pairwise comparison between end-to-end and cascaded captions across correctness, coverage, and naturalness dimensions, showing the proportions of better, tied, and worse cases under Gemini 3 Pro evaluation.

5.2 CLSP as a Speech Task Evaluator

5.2.1 Implementation Details

CLSP has 724M parameters in total, with 599M from SPEAR-XLarge and 125M from RoBERTa-base. Both training stages use 8 NVIDIA A100 80GB GPUs, with a batch duration of 800 seconds per GPU. The first stage is trained for 1.2M steps, and the second stage is fine-tuned for an additional 4k steps. We use the ScaledAdam (Yao et al., 2024) optimizer and the Eden (Yao et al., 2024) scheduler, with a peak learning rate of 0.045 for the first stage and 0.001 for the second stage.

5.2.2 Speech–Text Retrieval: Global and Fine-Grained

Evaluation Setup We construct our test set using 241 audio clips from the official ParaSpeechCaps test split, with durations ranging from 1 to 30 seconds, ensuring no overlap between training and test sets. Global captions are derived from the PSC-Base holdout split, and fine-grained captions are generated by our end-to-end pipeline.

Table 3: Global speech–text retrieval and fine-grained speech–text retrieval results. Baselines are evaluated using public checkpoints. Best results are in **bold**.

System	Speech-to-Text				Text-to-Speech			
	R@1	R@5	R@10	mAP@10	R@1	R@5	R@10	mAP@10
Global Speech–Text Retrieval								
LAION-AI CLAP (Wu et al., 2023)	0.4	2.5	5.4	1.5	0.8	3.3	5.0	1.9
GLAP (Dinkel et al., 2025)	1.7	5.0	9.5	3.4	1.7	5.8	10.8	3.9
ParaCLAP (Jing et al., 2024)	2.1	4.6	9.1	3.5	0.4	5.0	7.9	2.3
CLSP	**45.6**	**75.9**	**84.2**	**58.7**	**40.3**	**74.3**	**82.6**	**54.5**
Fine-Grained Speech–Text Retrieval								
LAION-AI CLAP (Wu et al., 2023)	0.4	3.3	5.4	1.6	1.2	2.5	2.9	1.7
GLAP (Dinkel et al., 2025)	4.6	14.5	21.2	9.1	2.1	7.5	13.3	4.4
ParaCLAP (Jing et al., 2024)	1.2	7.9	12.5	4.1	1.2	5.8	10.0	3.3
CLSP	**68.1**	**90.9**	**95.9**	**77.9**	**67.2**	**90.9**	**96.3**	**77.2**

Table 4: Zero-shot classification results, reported as WA (%) / UA (%). † indicates results taken from prior work; the others are evaluated using public checkpoints. Best results are in **bold**.

System	Emotion	Emotion	Emotion	Gender	Age
	IEMOCAP	RAVDESS	CREMA-D	RAVDESS	CREMA-D
LAION-AI CLAP (Wu et al., 2023)	32.6 / 29.1	14.4 / 13.5	21.0 / 19.2	58.6 / 58.6	1.3 / 1.8
GLAP (Dinkel et al., 2025)	32.5 / 27.0	7.8 / 13.0	13.6 / 17.9	72.6 / 72.6	27.1 / 32.6
ParaCLAP (Jing et al., 2024)	46.1 / 46.5	28.1 / 30.3	29.8 / 29.6	99.2 / 99.2	31.2 / 32.0
Auden-Voice CLAP (Huo et al., 2025)†	– / –	32.4 / –	30.2 / –	95.6 / –	38.5 / –
CLSP	**57.2 / 56.1**	**46.8 / 46.0**	**35.1 / 37.2**	**100.0 / 100.0**	**40.6 / 44.5**

Evaluation Metrics We use standard metrics, including Recall at rank 1/5/10 (R@1, R@5, R@10) and mean Average Precision at 10 (mAP@10). R@k measures the proportion of queries whose correct match appears in the top-k, while mAP@10 measures ranking quality within the top 10.

Baselines As the proposed tasks are novel, we evaluate representative open-source audio–text retrieval models, including LAION-AI CLAP, GLAP, and ParaCLAP, which are pre-trained on coarse-grained or task-specific supervision without fine-grained captions supervision. These models therefore serve to quantify the capability gap between existing approaches and our fine-grained modeling framework. Detailed descriptions of the baseline systems are provided in Appendix A.

Results Table 3 reports retrieval performance on both global and fine-grained tasks. Across all settings, CLSP consistently outperforms all baselines by a large margin across all evaluation metrics for both speech-to-text and text-to-speech retrieval, while the baselines perform close to random guessing. This substantial margin highlights the capability gap between existing models and CLSP, and underscores the effectiveness of training with global and fine-grained captions from FCaps.

5.2.3 Zero-shot Paralinguistic Classification

Evaluation Setup This task evaluates CLSP's ability to recognize paralinguistic attributes under diverse attribute sets, without any task-specific training. We focus on three representative paralinguistic dimensions: *emotion*, *gender*, and *age*. For each dimension, zero-shot classification is performed by computing the cosine similarity between a speech embedding and a set of natural-language text prompts describing candidates (e.g., "A speaker in a happy tone.", "A male speaker.", and "A middle-aged speaker."), and selecting the prompt with the highest similarity.

Our evaluations contain the following datasets:

- **IEMOCAP** (Busso et al., 2008): We use the 4-class emotion setup (happy/excited, angry, sad, neutral) with 5,531 utterances.
- **RAVDESS** (Livingstone and Russo, 2018): We use the speech part with 1,440 utterances, 8-class emotion categories (calm, happy, sad, angry, fearful, surprise, disgust, and neutral), and gender.
- **CREMA-D** (Cao et al., 2014): 7,442 utterances from 91 actors, with 6-class emotion categories (happy, sad, angry, fear, disgust, neutral) and speaker age, which we grouped into four bins (child, young adult, middle-aged, older).

Table 5: The correlation between the model-derived similarity scores and subjective MOS. Reported as Pearson / Spearman / Kendall's Tau correlation coefficients (r, ρ, τ). All results are statistically significant with $p < 0.01$.

System	Intrinsic Traits	Situational Traits	Fusion
LAION-AI CLAP (Wu et al., 2023)	0.679 / 0.664 / 0.467	0.194 / 0.184 / 0.126	0.588 / 0.597 / 0.425
GLAP (Dinkel et al., 2025)	0.372 / 0.340 / 0.234	0.169 / 0.138 / 0.102	0.350 / 0.344 / 0.241
ParaCLAP (Jing et al., 2024)	0.663 / 0.634 / 0.445	0.323 / 0.330 / 0.232	0.626 / 0.599 / 0.417
CLSP	**0.893 / 0.858 / 0.668**	**0.903 / 0.878 / 0.694**	**0.886 / 0.858 / 0.670**

Evaluation Metrics We report weighted accuracy (WA) and unweighted accuracy (UA), measuring overall and mean class accuracy, respectively.

Baselines We compare CLSP with strong baselines, including LAION-AI CLAP, GLAP, ParaCLAP, and Auden-Voice CLAP. For Auden-Voice CLAP, we use the ASR-pretrained variant without any supervised training on IEMOCAP, RAVDESS, or CREMA-D, ensuring strictly zero-shot evaluation. Baseline details are provided in Appendix A.

Results Table 4 reports zero-shot paralinguistic classification results on emotion, gender, and age recognition. Overall, CLSP consistently outperforms all baselines across datasets and paralinguistic dimensions. These results indicate that CLSP learns generalizable speech representations that capture diverse paralinguistic semantics without task-specific supervision.

5.2.4 Speech Style Similarity Scoring with Human Correlation

Subjective Evaluation Setup To validate the reliability of CLSP as an automated metric for assessing speech-text alignment, we investigate its consistency with human perception for three distinct paralinguistic dimensions. We conduct a meta-evaluation using the holdout split of ParaSpeechCaps. The evaluation focuses on *Intrinsic Traits*, *Situational Traits*, and their *Fusion*.

For human annotations, we recruited 20 experts with research backgrounds in speech processing to rate the matching degree between the audio and its corresponding text caption on a continuous scale. We then calculate the correlation between these subjective scores and model-predicted similarity scores across several baselines, using three standard statistical coefficients: the Pearson Correlation Coefficient (r), Spearman's Rank Correlation (ρ), and Kendall's Tau (τ). Details of the subjective evaluation are provided in Appendix E.

Baselines We continue to use open-source representative audio-text models, including LAION-AI CLAP, GLAP, and ParaCLAP as baseline systems.

Results The experimental results, as summarized in Table 5, demonstrate that CLSP significantly outperforms all baseline models in mirroring human perception across both intrinsic and situational traits, as well as their fusion. Furthermore, the consistent performance across all evaluated metrics ($r, \rho,$ and τ) confirms that CLSP not only tracks the absolute semantic matching quality linearly but also effectively preserves the relative ordinal ranking of samples in a manner that resonates with human expert judgment. In conclusion, these results validate CLSP as a robust and high-fidelity automated proxy for human subjective evaluation in paralinguistic audio-text matching tasks. For a more intuitive assessment, visual comparisons highlighting the strong concordance between model-predicted similarity scores and human subjective ratings are provided in Appendix F.

5.2.5 Ablation Study

To systematically analyze the effects of key design choices in training CLSP, we conduct ablation studies on multi-stage training, the weight λ in the multi-positive InfoNCE loss used in Stage Two, and task scheduling strategies for Stage Two. Detailed results are provided in Appendix G.

6 Conclusion

In this paper, we introduce FCaps, the largest speech dataset paired with fine-grained free-text speaking-style descriptions, together with CLSP, which learns fine-grained speech–text representations across multiple stylistic granularities. Extensive experiments demonstrate that CLSP performs reliably across a range of tasks, including global and fine-grained speech–text retrieval, zero-shot paralinguistic classification, and speech style similarity scoring that aligns strongly with human judgments. We hope this work encourages a shift in speaking style modeling from predefined, discrete attributes toward open-vocabulary, cross-granular, and speech-grounded natural language descriptions, facilitating more flexible and general-purpose speech–language representations.

Limitations

While our approach demonstrates strong performance across a range of tasks, it has several limitations. First, CLSP relies on a pre-trained speech and audio unified encoder that is trained only on English speech, and therefore our model is limited to English. Second, publicly available paralinguistic speech data with diverse style attributes remains limited, particularly in its coverage of underrepresented accents, emotional expressions, and expressive speaking styles. We leave extensions toward more diverse and multilingual speech representations as an important direction for future work.

Ethics Statement

All data used in this work were collected and processed in accordance with relevant ethical guidelines and licensing terms. The speech samples are sourced from publicly available datasets. We will only release the annotated captions and associated metadata that reference the original audio, and will not redistribute the source audio files. Human annotations were conducted by trained annotators, who were fairly compensated.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, and 1 others. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359.

Houwei Cao, David G. Cooper, Michael K. Keutmann, and 1 others. 2014. CREMA-D: crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affect. Comput.*, 5(4):377–390.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Heinrich Dinkel, Zhiyong Yan, Tianzi Wang, and 1 others. 2025. GLAP: general contrastive audio-text pretraining across domains and languages. *Preprint*, arXiv:2506.11350.

Anuj Diwan, Zhisheng Zheng, David Harwath, and Eunsol Choi. 2025. Scaling rich style-prompted text-to-speech datasets. In *Proc. EMNLP*, Suzhou.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP learning audio concepts from natural language supervision. In *Proc. ICASSP*, Rhodes Island.

Zhifang Guo, Yichong Leng, Yihan Wu, and 1 others. 2023. PromptTTS: controllable text-to-speech with text descriptions. In *Proc. ICASSP*, Rhodes Island.

Jiarui Hai, Karan Thakkar, Helin Wang, and 1 others. 2024. DreamVoice: Text-guided voice conversion. In *Proc. Interspeech*, Kos Island.

Haorui He, Zengqiang Shang, Chaoren Wang, and 1 others. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *Proc. SLT*, Macao.

Kaiming He, Haoqi Fan, Yuxin Wu, and 1 others. 2020. Momentum contrast for unsupervised visual representation learning. In *Proc. CVPR*, Seattle.

Mingyue Huo, Wei-Cheng Tseng, Yiwen Shao, and 1 others. 2025. Auden-Voice: General-purpose voice encoder for speech and language understanding. *Preprint*, arXiv:2511.15145.

Shengpeng Ji, Qian Chen, Wen Wang, and 1 others. 2025. ControlSpeech: Towards simultaneous and independent zero-shot speaker cloning and zero-shot language style control. In *Proc. ACL*, Vienna.

Shengpeng Ji, Jialong Zuo, Minghui Fang, and 1 others. 2024. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *Proc. ICASSP*, Seoul.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, and 1 others. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zeyu Jin, Jia Jia, Qixin Wang, and 1 others. 2024. SpeechCraft: A fine-grained expressive speech dataset with natural language description. In *Proc. ACM MM*, Melbourne.

Xin Jing, Andreas Triantafyllopoulos, and Björn Schuller. 2024. ParaCLAP – towards a general language-audio model for computational paralinguistic tasks. In *Proc. Interspeech*, Kos Island.

Masaya Kawamura, Ryuichi Yamamoto, Yuma Shirahata, and 1 others. 2024. LibriTTS-P: A corpus with speaking style and speaker identity prompts for text-to-speech and style captioning. In *Proc. Interspeech*, Kos Island.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

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

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, and 1 others. 2025. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. *Preprint*, arXiv:2510.12720.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, and 1 others. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Proc. ACL*, Bangkok.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A large-scale speaker identification dataset. In *Proc. Interspeech*, Stockholm.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, and 1 others. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. In *Proc. Interspeech*, Dublin.

OpenAI. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

Alec Radford, Jong Wook Kim, Chris Hallacy, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Julius Richter, Yi-Chiao Wu, Steven Krenn, and 1 others. 2024. EARS: an anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. In *Proc. Interspeech*, Kos Island.

Haoqin Sun, Jingguang Tian, Jiaming Zhou, and 1 others. 2025. RA-CLAP: relation-augmented emotional speaking style contrastive language-audio pretraining for speech retrieval. In *Proc. Interspeech*, Rotterdam.

Helin Wang, Jiarui Hai, Dading Chong, and 1 others. 2025. CapSpeech: enabling downstream applications in style-captioned text-to-speech. *Preprint*, arXiv:2506.02863.

Yusong Wu, Ke Chen, Tianyu Zhang, and 1 others. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proc. ICASSP*, Rhodes.

Guanrou Yang, Chen Yang, Qian Chen, and 1 others. 2025a. EmoVoice: LLM-based emotional text-to-speech model with freestyle text prompting. In *Proc. ACM MM*, Dublin.

Xiaoyu Yang, Yifan Yang, Zengrui Jin, and 1 others. 2025b. SPEAR: A unified SSL framework for learning speech and audio representations. *Preprint*, arXiv:2510.25955.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, and 1 others. 2024. Zipformer: A faster and better encoder for automatic speech recognition. In *Proc. ICLR*, Vienna.

A Baseline Details

- **LAION-AI CLAP** (Wu et al., 2023): An open-source[5] CLIP-style dual-encoder model with 158M parameters. The audio encoder uses pre-trained HTSAT-tiny, with audio representations extracted from the penultimate layer before projection. The text encoder uses RoBERTa-base, with text representation taken from the final-layer [CLS] token. LAION-AI CLAP is trained on large-scale audio-text paired data, including AudioCaps, Clotho, and the LAION-Audio-630K dataset. Training uses bidirectional contrastive learning with a symmetric cross-entropy loss.
- **GLAP** (Dinkel et al., 2025): An open-source[6] dual-encoder CLAP-style model with 855M parameters that aims to learn unified audio-text representations across speech, sound, and music domains. GLAP employs a pre-trained general-purpose audio encoder and a multilingual text encoder based on Sonar for text representations. GLAP is trained on large-scale audio-text paired data spanning speech (439.4M, 411k hours), sound (5.9M, 23.8k hours), and music (3k, 19.3 hours). Training uses a bidirectional contrastive objective based on a sigmoid loss.
- **ParaCLAP** (Jing et al., 2024): An open-source[7] dual-encoder CLAP-style model with 276M parameters. The speech encoder is initialized from a pruned version of Wav2Vec2-Large-Robust that has been further fine-tuned for dimensional speech emotion recognition on MSP-Podcast. Audio representations are obtained by pooling the hidden states of the final Transformer layer. The text encoder is bert-base-uncased, and the text representation is taken from the final-layer [CLS] token. ParaCLAP is trained on the MSP-Podcast dataset, which contains nine emotion categories, utilizing bidirectional contrastive learning with a symmetric cross-entropy loss.
- **Auden-Voice CLAP** (Huo et al., 2025): A dual-encoder CLAP-style model with 281M parameters, which remains closed-source at the time of submission. The speech encoder is initialized from an ASR-pretrained Zipformer-L encoder trained on large-scale in-house Chinese ASR data using the RNNT loss. Audio representations are obtained by mean-pooling the hidden

[5] https://github.com/LAION-AI/CLAP

[6] https://github.com/xiaomi-research/dasheng-glap

[7] https://github.com/KeiKinn/ParaCLAP

states of the final Transformer layer. The text encoder is RoBERTa-base, and text representations are taken from the final-layer [CLS] token. Different continual pre-training settings are applied starting from the ASR-pretrained encoder, including speaker identification trained on VoxCeleb2 (974k utterances and 2,026 hours), paralinguistic attribute prediction trained on CREMA-D, RAVDESS, IEMOCAP, and TESS (18.3k utterances and 20 hours), and a multi-task setting combining speaker identification and paralinguistic tasks. Encoders obtained from each setting are used to initialize different CLAP variants, which are further trained on ParaSpeechCaps using both PSC-Base and PSC-Scaled splits (1.036M audio-text pairs, 2,700 hours) with bidirectional speech–text contrastive learning.

B Case Study: User Prompt Composition for Fine-Grained Caption Generation

This case study qualitatively examines how different user prompt compositions influence the fine-grained captions generated by Qwen3-Omni-30B-A3B-Captioner.

Setup We compare three prompt compositions:

- **Audio-only**: The captioner takes audio as input and generates captions in its default manner.
- **Audio + User Prompt (w/o Tags)**: A user prompt (Figure 7) is used to discourage spoken-content transcription, environment-related descriptions, and audio quality assessments.
- **Audio + User Prompt (w/ Tags)**: A user prompt with human-annotated attributes (Figure 8) is used to guide caption generation toward accurate realization of specified attributes, especially those that are rare or inherently ambiguous when inferred from audio alone.

Analysis As shown in Table 6, different prompt compositions lead to distinct captioning behaviors.

- When no user prompt is applied, the generated caption includes spoken-content transcription, detailed descriptions of environmental background noise, audio quality assessments, speculative mentions of editing artifacts and electronic tones, as well as explicit declarations about the absence of certain elements, all highlighted in red. Such speaker-independent and verbose content can distract contrastive learning from speaker-centric representations, encouraging reliance on shallow lexical cues from verbatim transcription or recording-specific artifacts.
- Introducing a lightweight user prompt results in a more concise and speaker-focused caption. However, the inferred accent remains coarse, highlighted in red, reflecting the intrinsic ambiguity of accent identification from short speech segments, where many regional accents exhibit highly similar acoustic patterns and are difficult to reliably distinguish based on audio alone.
- When human-annotated attributes are additionally provided, the specified accent can be realized accurately in the generated caption, highlighted in green. The resulting description not only reflects the correct accent but also maintains a coherent temporal structure that captures changes in speaking style and delivery.

Overall, this comparison illustrates how human annotations can resolve ambiguity in inherently hard-to-delineate speaker traits, while preserving the narrative nature of fine-grained captions.

C Case Study: Multi-Positive Caption Generation

Analysis Table 7 presents multiple fine-grained captions generated for the same speech clip using different random seeds. All captions consistently capture core speaker attributes, including gender, accent, and broadly similar pitch range and speaking rate, indicating that they remain well grounded in the same underlying speech signal. At the same time, the captions exhibit notable variation in lexical choice, descriptive emphasis, and narrative structure. Some emphasize overall delivery style and speaker demeanor, while others describe within-utterance dynamics such as changes in pace, pauses, or shifts in intonation. Together, these captions constitute multiple semantically non-identical textual views of the same speech.

D Detailed Protocol of LLM-as-Judges

The LLM-as-Judges evaluation follows three criteria, each assessed using a five-level rubric that specifies both failure modes and ideal behaviors, as detailed in Figure 9.

E Details of Subjective Evaluation

Evaluation Data Selection To validate whether CLSP scores align with human perception, we con-

Table 6: Case study of different user prompt compositions and corresponding generated fine-grained captions.

User Prompt Composition	Fine-grained Caption
Audio-only	The audio clip begins abruptly, mid-sentence, as a male speaker with a distinct Afro-Caribbean accent continues a thought: "…of of the idea that has been the same idea for a thousand years that they believe that—". His delivery is measured and analytical, marked by a slight hesitation and repetition at the start, indicating an unscripted, conversational tone. The speaker's voice is steady, low-to-mid in pitch, and slightly gravelly, with clear articulation and no signs of emotional distress or heightened affect. The environment is quiet, featuring only a faint electronic hum and minimal room reverberation, pointing to a small, acoustically treated indoor setting such as a home studio or office. The recording is of high technical quality, with no distortion or background interference, and the speaker's voice is captured close to the microphone, ensuring clarity and intelligibility. At the moment the speaker is about to elaborate further, the audio is abruptly cut off, truncating his final word and leaving the thought incomplete. This sudden interruption is accompanied by a sharp, digital click, characteristic of a hard edit or technical error, and is immediately followed by a brief, high-frequency electronic tone. The tone is static and unchanging, likely the result of a digital artifact from the editing process, and is not part of the original speech. No other sounds—such as music, ambient noise, or environmental cues—are present, and the absence of audience reactions or other voices indicates a solitary recording context. In summary, the clip features a single, articulate male speaker with a Caribbean accent discussing a persistent, long-standing idea in an unembellished, contemplative manner. The audio is clear and focused, set in a quiet, private environment, and is abruptly interrupted by a technical edit, which introduces a digital click and an electronic tone, cutting off the speaker mid-sentence and leaving the discussion unresolved.
Audio + User Prompt w/o Tags	A male speaker with a medium-low pitched, slightly gravelly voice and a non-native English accent delivers a slow, deliberate, and thoughtful statement. His speech is marked by a slight repetition at the beginning and pauses between phrases, with a steady, calm, and measured pace. The intonation is even and declarative, conveying a sense of reflection and certainty, as if presenting a considered opinion or fact.
Audio + User Prompt w/ Tags	A mature male with a clear, medium-low pitched voice and a Jamaican accent speaks in a measured, deliberate manner. He begins with a slightly rushed, emphatic tone and a moderate pitch, conveying a sense of conviction. After a short pause, he continues at a steady, thoughtful pace with a slightly lower pitch and a more subdued, reflective tone. The speech is delivered at a moderate volume with a controlled, conversational delivery.

Table 7: Case study of multi-positive caption generation for the same speech clip using different random seeds.

Seed	Fine-grained Caption
1	A middle-aged male with a clear, medium-pitched voice and an Australian accent speaks in a conversational, explanatory tone. His speech is delivered at a fast, steady pace, with a moderate volume and smooth, fluid articulation. The intonation is relatively flat and consistent, conveying a matter-of-fact, informative style.
2	A male speaker with a medium-low pitched, resonant voice and an Australian accent speaks in a measured, deliberate manner. His delivery is calm and conversational, with a steady rhythm and even pacing, using a moderate volume and a flat, unemotional tone.
3	A mature male with a clear, medium-pitched voice and an Australian accent speaks in a conversational, informal style. He begins at a fast, energetic pace with an emphatic, high pitch and moderately loud volume, using a casual, slightly breathy tone. After a brief pause, he continues at a steady, conversational pace with a moderate pitch and volume, his tone remaining expressive and fluctuating throughout the fluent delivery.
4	A mature male with a medium-low pitched, resonant voice and a distinct Australian accent speaks in a measured, confident manner. His delivery is steady and deliberate, with a moderate speaking rate and clear articulation. The tone is calm and conversational, suggesting an explanatory or persuasive intent. He maintains a consistent volume and pitch, with subtle inflections that convey a sense of authority and familiarity.
5	A mature male with a clear, medium-pitched voice and an Australian accent speaks in a conversational, explanatory style, similar to a commentator or analyst. He begins with a moderate pace and clear enunciation, using a mid-to-low pitch that rises slightly for emphasis. After a short pause, he continues at a slightly faster pace with a more declarative tone, and his pitch drops again as he concludes with a downward inflection, maintaining a steady, conversational delivery throughout.

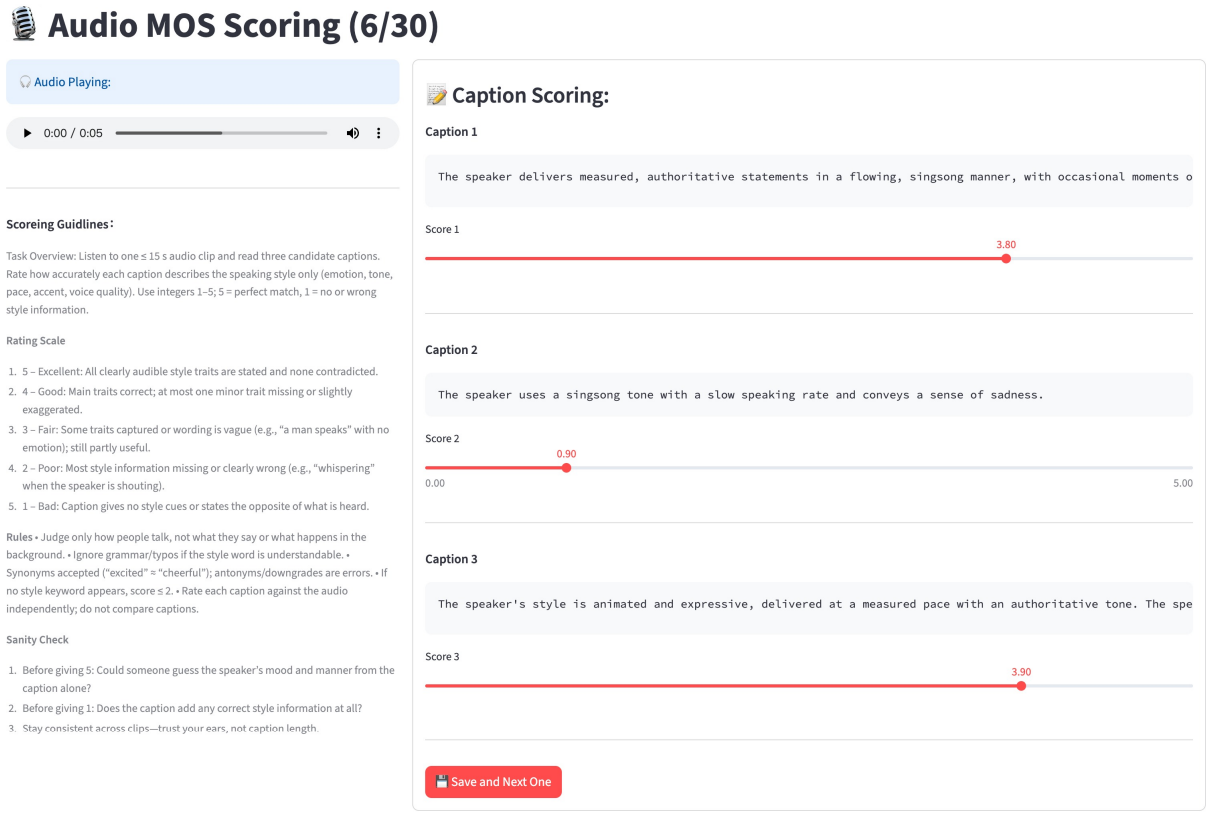

Figure 5: Annotation UI for raters to annotate the alignment score between one audio and several candidate captions.

duct a meta-evaluation using the evaluation set of ParaSpeechCaps. This dataset contains diverse audio samples paired with style descriptions covering two key paralinguistic dimensions:

- Intrinsic Speaker Traits: Speaker characteristics tied to a speaker's identity, such as gender, age, timbre/texture, pitch, accent, and so on.
- Situational Traits: Dynamic aspects including speaking rate, emotion, expressivity, volume, and other speaker style-related attributes.
- Fusion: Complex audio captions that combine both intrinsic and situational descriptors.

For each category, we randomly select 30 audio clips for human evaluation and scoring, and adopt a text-based large language model Qwen3-30B-A3B-Instruct-2507[8] to rewrite and manually filter the official captions provided, ensuring that the final captions only contain the intrinsic speaker or situational traits for evaluation.

Human Annotation Protocol Figure 5 illustrates the user interface used for human subjective evaluation. For each evaluation instance, 20 raters were presented with an audio clip together with three candidate captions describing the intrinsic speaker or situational traits. Raters were allowed to replay the audio as needed before providing their judgments. The correlation score was assessed using a Mean Opinion Score (MOS) protocol. Raters assigned a score to each caption via a slider on a continuous scale from 0.0 to 5.0, where higher values indicate greater perceptual similarity and stronger consistency between the caption and the audio. Detailed scoring guidelines were displayed alongside the interface. Raters were instructed to evaluate each caption independently based on how accurately it captured style attributes of the speech, while ignoring irrelevant factors such as grammatical errors or minor wording variations.

Evaluation Metrics We employ three widely recognized statistical coefficients to measure the agreement between our model-derived similarity scores and human subjective ratings:

- Pearson Correlation Coefficient (r): Evaluates the linear relationship between the model predictions and human scores.
- Spearman's Rank Correlation (ρ): Assesses the monotonic relationship, reflecting how well the

[8] https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507

model preserves the relative ranking of samples.

- Kendall's Tau (τ): Measures the ordinal association and pairwise agreement between the two sets of rankings, providing a robust check for consistency.

F Visualization of Subjective Evaluation

Figure 6 provides a visual comparison of the correlation between model-predicted similarity scores and human subjective ratings across four models and three paralinguistic trait categories. Compared to baseline models, the scatter plots for CLSP (bottom row) exhibit a significantly tighter clustering of data points around the red linear regression lines. This pattern is consistent across all sub-categories. The higher density of points along the diagonal confirms that CLSP can reliably replicate the nuanced judgments of human experts.

The visual trends observed in Figure 6, together with the strong Pearson, Spearman, and Kendall's Tau correlations reported in Table 5, suggest that CLSP provides a reliable automated proxy for assessing speech-text style alignment. In particular, CLSP captures relative ranking and preference patterns that are consistent with human judgments, supporting its use for large-scale evaluation in speech-text matching scenarios where human annotation is costly.

G Ablation Studies

We conduct ablation studies on the following components to provide an in-depth understanding of the training of CLSP:

- **Multi-stage training**: We compare models trained with different stages (see Appendix G.1).
- **Target weight**: We compare different target weights λ in multi-positive InfoNCE loss used in Stage Two (see Appendix G.2).
- **Task scheduler**: We compare different task scheduling strategies used in Stage Two (see Appendix G.3).

G.1 Ablation Study: Multi-Stage Training

Table 8 presents an ablation study examining the effectiveness of each training stage. Training with only Stage 1 primarily benefits fine-grained speech–text retrieval, whereas training with only Stage 2 achieves only limited performance on both global and fine-grained retrieval tasks. Combining both stages yields the best performance across all metrics, demonstrating that each stage contributes effectively to both global and fine-grained speech–text alignment.

G.2 Ablation Study: Target Weight

Table 9 reports an ablation study on the loss weight λ. Performance remains stable across a broad range of λ values, indicating that the proposed training objective is not overly sensitive to the exact choice of the target weight. Among the tested settings, $\lambda = 0.5$ achieves the best overall performance on both global and fine-grained speech–text alignment, and is therefore used in all experiments.

G.3 Ablation Study: Task Scheduler

Table 10 presents an ablation study of different task scheduling strategies.

For static mixtures, decreasing the sampling probability of Task 1 shifts the training focus towards fine-grained discrimination via semantic consistency, leading to consistent improvements in fine-grained retrieval performance but a notable degradation in global retrieval, particularly when p_0 falls below 0.2. Conversely, assigning a higher sampling probability to Task 1 improves both global and fine-grained performance by encouraging effective cross-granularity generalization, although fine-grained retrieval does not reach the level achieved when training exclusively with Task 2. These reveal a trade-off in static scheduling strategies.

In contrast, dynamic schedulers effectively combine the advantages of both tasks and achieve overall stronger and more balanced performance across global and fine-grained retrieval. The best results are obtained by a dynamic scheduler that gradually shifts the sampling distribution from Task 1 to Task 2, with $p_0 = 0.95$, $p_{\min} = 0.50$, and $T = 10{,}000$, confirming the effectiveness of curriculum-style task scheduling.

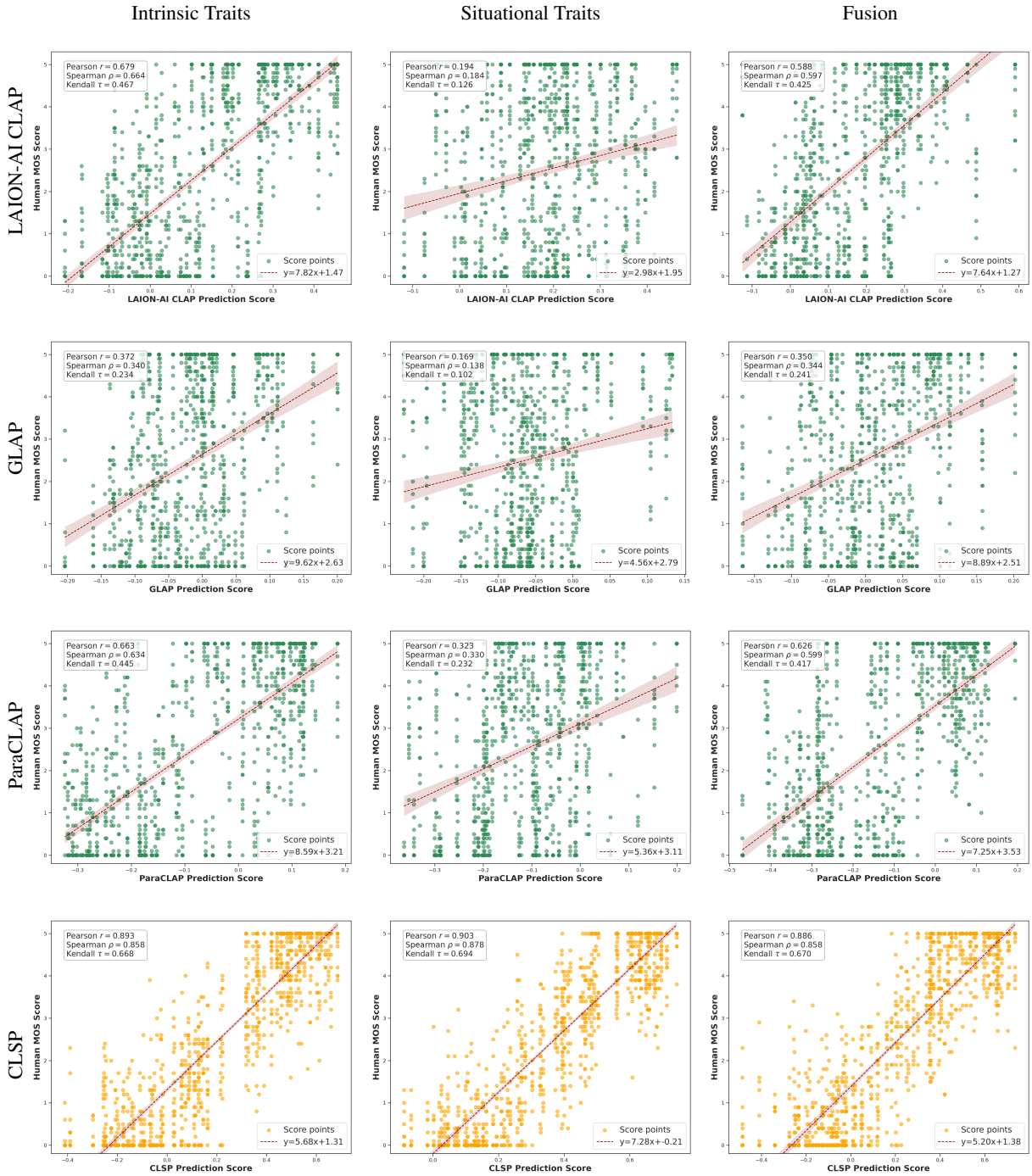

Figure 6: Correlation analysis between model-predicted similarity scores and subjective human ratings across different models and trait categories (Intrinsic, Situational, and Fusion). The green scatter plots demonstrate the alignment between automated metrics and human perception using the ParaSpeechCaps evaluation set. The dashed red lines represent the linear regression fit.

Table 8: Ablation study of different training stages, evaluated on global speech–text retrieval and fine-grained speech–text retrieval tasks. ✓ / ✗ indicate whether a training stage is performed.

Stage 1	Stage 2	Global Speech–Text Retrieval		Fine-Grained Speech–Text Retrieval		Avg.
		Speech-to-Text mAP@10	Text-to-Speech mAP@10	Speech-to-Text mAP@10	Text-to-Speech mAP@10	
✓	✗	11.1	9.6	69.6	64.0	38.6
✗	✓	38.3	38.5	35.3	32.6	36.2
✓	✓	**58.7**	**54.5**	**77.9**	**77.2**	**67.1**

Table 9: Ablation study of different loss weights, evaluated on global speech–text retrieval and fine-grained speech–text retrieval tasks.

λ	Global Speech–Text Retrieval		Fine-Grained Speech–Text Retrieval		Avg.
	Speech-to-Text mAP@10	Text-to-Speech mAP@10	Speech-to-Text mAP@10	Text-to-Speech mAP@10	
0.3	53.4	52.6	**77.9**	77.0	65.2
0.4	56.0	**55.1**	77.5	75.5	66.0
0.5	**58.7**	54.5	**77.9**	**77.2**	**67.1**
0.6	56.4	55.0	77.7	76.9	66.5
0.7	57.8	54.4	77.3	75.6	66.3

Table 10: Ablation study of different task schedulers, evaluated on global speech–text retrieval and fine-grained speech–text retrieval tasks.

p_0	$p_{\min}$	T	Global Speech–Text Retrieval		Fine-Grained Speech–Text Retrieval		Avg.
			Speech-to-Text mAP@10	Text-to-Speech mAP@10	Speech-to-Text mAP@10	Text-to-Speech mAP@10	
Static Mixture							
0.00	–	–	20.4	19.0	84.9	79.6	51.0
0.20	–	–	42.7	41.1	83.9	78.9	61.7
0.30	–	–	51.6	49.2	81.3	76.4	64.6
0.40	–	–	51.1	50.6	79.2	77.1	64.5
0.50	–	–	52.1	51.8	79.6	77.4	65.2
0.60	–	–	54.4	52.5	79.7	76.7	65.8
0.70	–	–	55.8	52.9	78.8	77.9	66.4
0.80	–	–	56.9	54.5	77.7	76.0	66.3
0.90	–	–	57.6	56.5	77.5	75.7	66.8
1.00	–	–	57.6	55.9	77.9	74.9	66.6
Dynamic Mixture							
0.95	0.05	5000	53.5	53.1	79.9	75.5	65.5
0.95	0.50	5000	55.7	54.4	78.0	76.9	66.3
0.95	0.50	10000	**58.7**	54.5	77.9	77.2	**67.1**
0.95	0.50	15000	55.6	55.5	78.2	76.1	66.4

Detailed Captioner User Prompt

Your task is to generate a caption describing ***only the characteristics of the speaker's voice.***

Audio: *{audio}*

CRITICAL RULES
1. ***NEVER*** *describe the content of the speech. Do not quote any words or phrases.* ***NEVER*** *contain quotation marks ("").*
2. ***FOCUS ONLY ON THE HUMAN VOICE. NEVER*** *describe background, environment, audio quality.*
3. ***NEVER*** *mention the absence of characteristics (describe only what is present, not mention what is not present).*
4. ***NEVER*** *over-interpret or guess.*
5. Failure to follow these rules will result in an invalid output.

Good Example
A young male with a clear, medium-high pitched voice and an American accent speaks in a casual, conversational style, much like a reviewer or vlogger. He begins at a fast, rushed pace with a highly energetic and emphatic intonation, using a high pitch to express strong emphasis. After a slight inhale, he continues to speak quickly and enthusiastically, maintaining a moderately loud volume and an expressive, fluctuating tone throughout the fluent delivery.

YOUR CAPTION:

Figure 7: User prompt for detailed captioner.

Detailed Captioner User Prompt w/ Tags

Your task is to generate a caption describing ***only the characteristics of the speaker's voice.***

Audio: *{audio}*

Use the following tags in the caption:
- ***Accent****: {accent}*
- ***Speaking Rate****: {speaking_rate}*
- ***Emotion / Expressiveness****: {situational_tags}*

CRITICAL RULES
1. ***NEVER*** *describe the content of the speech. Do not quote any words or phrases.* ***NEVER*** *contain quotation marks ("").*
2. ***FOCUS ONLY ON THE HUMAN VOICE. NEVER*** *describe background, environment, audio quality.*
3. ***NEVER*** *mention the absence of characteristics (describe only what is present, not mention what is not present).*
4. ***NEVER*** *over-interpret or guess.*
5. Failure to follow these rules will result in an invalid output.

Good Example
A young male with a clear, medium-high pitched voice and an American accent speaks in a casual, conversational style, much like a reviewer or vlogger. He begins at a fast, rushed pace with a highly energetic and emphatic intonation, using a high pitch to express strong emphasis. After a slight inhale, he continues to speak quickly and enthusiastically, maintaining a moderately loud volume and an expressive, fluctuating tone throughout the fluent delivery.

YOUR CAPTION:

Figure 8: User prompt with human-annotated tags for detailed captioner.

Caption Quality Assessment API Prompt

You are an evaluator assessing speech style caption quality.

Audio: *{audio}*

Caption A: *{caption_a}*

Caption B: *{caption_b}*

Task:
Evaluate two captions based on the audio content; do not rely on external information.
Assign scores from 1 to 5 for each caption on the three dimensions below, according to the scoring guidelines.
Provide a brief explanation comparing the two captions.

Scoring Guidelines:
- Audio-grounded correctness:
Does the caption accurately reflect what can be heard in the audio?
Penalize factual errors, misattributions (e.g., confusing the speaker's emotion with described content), or perceptual confusions (e.g., high volume vs. high pitch).
5: Fully accurate and faithful to the audio; no factual or perceptual errors.
4: Mostly accurate; minor imprecision that does not affect the main meaning.
3: Partially correct; noticeable errors or ambiguities, but some correct aspects remain.
2: Largely incorrect; multiple factual or perceptual errors that misrepresent the audio.
1: Completely incorrect or misleading; does not reflect the audio at all.

- Coverage:
Does the caption adequately capture salient speaking-style attributes that are evident in the audio?
Penalize missing important attributes, but do not penalize the absence of attributes that are not clearly audible.
5: Captures all salient attributes evident in the audio.
4: Captures most salient attributes; some minor attributes may be missing.
3: Captures some attributes but misses some important ones.
2: Captures very few attributes; most are missing.
1: Fails to capture any speaking-style attributes.

- Naturalness:
Is the caption fluent, grammatical, and natural-sounding as a human-written description?
Penalize awkward phrasing, grammatical errors, parenthetical explanations, or any patterns indicative of AI-generated text.
5: Fully fluent, grammatical, and natural-sounding; indistinguishable from a human-written caption.
4: Generally fluent and natural with minor awkward phrasing or mild artificiality.
3: Understandable but noticeably awkward or artificial in places.
2: Poor fluency or clearly artificial, with stylistic or structural patterns typical of AI-generated text.
1: Incoherent, ungrammatical, or strongly artificial and unnatural.

Output Format:
Respond in JSON format only. Do not output anything outside the JSON.

```
{
  "explanation": "<brief explaining>",
  "caption_a": {
    "correctness": <integer from 1 to 5>,
    "coverage": <integer from 1 to 5>,
    "naturalness": <integer from 1 to 5>
  },
  "caption_b": {
    "correctness": <integer from 1 to 5>,
    "coverage": <integer from 1 to 5>,
    "naturalness": <integer from 1 to 5>
  }
}
```

Figure 9: Detailed protocol of LLM-as-Judges.