SLS: STRUCTURED LOCAL SUPERVISION FOR FINE-GRAINED SPEECH-TEXT ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained speech-text alignment, such as understanding subtle shifts in speaking rate or vocal delivery, requires audio-text models to assign high similarity to accurate descriptions, while distinguishing them from inaccurate ones. We introduce Structured Local Supervision (SLS), a framework that adapts pretrained audio-text models using structured captions with SigLIP and supervised contrastive learning. Accurate captions encourage invariance to paraphrasing and partial descriptions, while inaccurate captions define targeted attribute contrasts and guide in-batch audio-negative mining. Across three backbone families, SLS substantially improves speech retrieval, classification, and controlled fine-grained ranking. Notably, SLS fine-grained alignment extends beyond retrieval, offering a lightweight, zero-shot proxy for instruction-conditioned text-to-speech evaluation. SLS rivals multi-billion-parameter audio LLMs in correlation with human adherence ratings. The findings highlight SLS as both a general recipe for attribute-sensitive representations and an efficient, scalable foundation for generative speech evaluation and alignment objectives.

1 INTRODUCTION

Large-scale contrastive pretraining has been extensively studied to align audio and natural-language descriptions in a shared high-dimensional embedding space for applications like retrieval and zero-shot classification (Wu et al., 2023; Elizalde et al., 2023). Conventional methods encourage greater similarity between positive audio-text pairs while decreasing similarity between unpaired data within the randomly sampled batch (negative pairs) (Oord et al., 2018; Zhai et al., 2023). Meanwhile, speech retrieval and instruction-conditioned text-to-speech (TTS) evaluation require audio-text representations to capture subtle audible differences. E.g., for a caption describing rapid speech, a faithful paraphrase like "quick" remains accurate, while changing to "slow" creates an inaccurate description. In the same example, a partial description that omits speaking rate but correctly describes other audible attributes remains accurate. However, conventional contrastive learning lacks direct supervision from multiple positives and informative negatives to teach models to distinguish these slight distinctions (Kamath et al., 2024; Ghosh et al., 2024). Hence, this paper focuses on the broader representation learning challenge of assigning high similarity to accurate descriptions at different levels of detail while distinguishing subtle changes that contradict the audio.

We propose Structured Local Supervision (SLS), a general audio-text model adaptation method for fine-grained speech-text alignment (Figure 1). We construct accurate and inaccurate captions from large paired audio-caption datasets to support audio-to-text local supervision. *Accurate captions* teach the model to assign high similarity to faithful paraphrases and incomplete but valid descriptions. *Inaccurate captions* contradict specific speech attributes and serve as queries for selecting existing in-batch audio negatives. E.g., a caption with "quickly" replaced by "slowly" is used to retrieve contrasting recordings for the text-to-audio local supervision. Moreover, SLS integrates SigLIP correspondence training with supervised contrastive learning (SupCon) (Zhai et al., 2023; Khosla et al., 2020), where SigLIP offers broad audio-text alignment, while SupCon emphasizes the structured relations between accurate and inaccurate pairs.

To show SLS's fine-grained speech understanding, we build a speech-centered evaluation that connects retrieval, classification, and ranking. Classification spans speaking style, recording conditions, conversation, and speech content. Ranking tests local discrimination between audio and text by

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

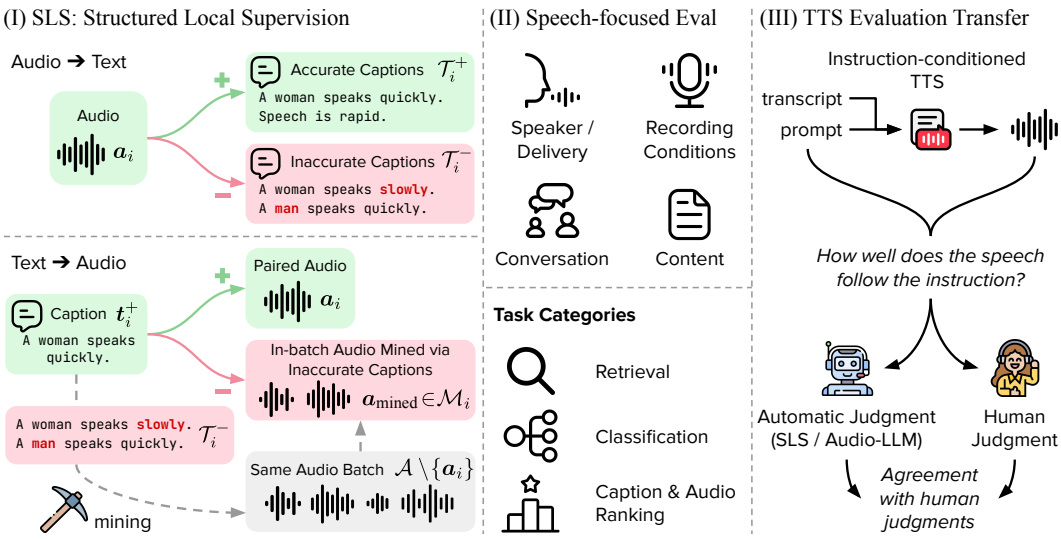


Figure 1: Overview of SLS. (I) SLS learns from constructed accurate and inaccurate captions and uses inaccurate captions to mine contrasting recordings from the training batch. The + and − arrows mark positive and negative pairs for SigLIP and SupCon. (II) Speech-focused evaluation covers retrieval, classification, and ranking. (III) Automatic TTS adherence scores are compared with listener ratings of how well reference and synthesized speech follow the instructions.

ranking accurate/inaccurate captions or recordings. Experiments show that SLS improves these tasks over three backbone families, including VoiceCLAP (LAION, 2026), M2D-CLAP (Niizumi et al., 2025), and PE_{AV} (Vyas et al., 2026). Qualitative analysis further shows stronger separation between accurate descriptions and targeted attribute contradictions after SLS adaptation.

Beyond standard applications, SLS speech-text alignment extends the utility of adapted audio-text models to generative speech evaluation by serving as lightweight, perceptually grounded evaluators. In instruction-conditioned TTS, evaluating whether synthesized speech adheres to stylistic directives typically requires expensive human listening studies or heavy audio large language models (LLMs) (Huang et al., 2025). We show that SLS-adapted embeddings bridge this gap by offering a zero-shot similarity metric that correlates with human ratings. Our contributions are:

1. **Method:** SLS connects constructed accurate and inaccurate captions with audio-negative mining, supplying relations for bidirectional local supervision to adapt existing audio-text models.
2. **Speech understanding:** Empirical evidence and analysis reveal that SLS improves speech retrieval, classification, and ranking across backbones.
3. **Transfer to TTS instruction-adherence evaluation:** We bridge representation learning with generative speech evaluation, showing that SLS-adapted models rival several audio LLMs in human-adherence correlation. The findings support audio-text encoders as practical, compute-efficient proxies for TTS evaluation, motivating future investigation as reward signals.

2 Structured Local Supervision

SLS adapts audio-text models through structured audio-caption relations (Section 2.1) and bidirectional local supervision (Section 2.2). Here, "structured" refers to explicit accurate and inaccurate audio-caption relations, and "local" refers to the targeted caption-defined contrasts.

2.1 Structured Caption Relations and Audio Mining

Consider a global batch of B audio embeddings $\mathcal{A} = \{\boldsymbol{a}_i\}_{i=1}^{B} \subset \mathbb{R}^D$ with anchor $\boldsymbol{a}_i$ having up to N_T^+ accurate captions $\mathcal{T}_i^+ \subset \mathbb{R}^D$ and up to N_T^- inaccurate captions $\mathcal{T}_i^- \subset \mathbb{R}^D$. Accurate captions teach the model to accept faithful paraphrases and partial but correct descriptions. Inaccurate captions offer targeted attribute contrasts by changing an attribute while preserving the surrounding

content. Moreover, $\mathcal{T}_i^+$ and $\mathcal{T}_i^-$ include instruction-style captions and hand-written single-attribute templates. Hence, the constructed captions offer local *audio-to-text* discrimination supervision. Appendix B.1 describes caption construction.

Supporting the *text-to-audio* supervision, we use inaccurate captions to mine existing in-batch recordings as training negatives. Related negative-sampling strategies have been studied for audio-text retrieval (Xie et al., 2023), but SLS uses explicitly constructed inaccurate captions as queries to mine in-batch audio negatives, linking audio-negative selection to targeted attribute contrasts. The miner selects top N_A^- recordings with the highest similarity to the anchor's inaccurate captions:

$$\mathcal{M}_i = \left\{ \boldsymbol{a}_j : j \in \arg\operatorname*{top\text{-}}_{j \in \mathcal{E}_i} N_A^- \frac{1}{\left|\mathcal{T}_i^-\right|} \sum_{\boldsymbol{t}_i^- \in \mathcal{T}_i^-} \mathrm{sg}(\boldsymbol{t}_i^-)^\intercal \mathrm{sg}(\boldsymbol{a}_j) \right\}, \tag{1}$$

where $\mathcal{E}_i$ is the set of indices that excludes the anchor and recordings sharing known positive captions. The stop-gradient operator sg makes explicit that $\operatorname{top}\text{-}N_A^-$ selection only determines candidate indices, while gradients flow through the scores in the training objective. Together, structured caption relations and mined audio negatives define the relations for bidirectional local supervision.

2.2 Training Objective

Correspondence supervision. SigLIP correspondence learning teaches the model to align paired audio and captions across the batch via sigmoid loss, while local supervision distinguishes accurate matches from targeted alternatives. SigLIP pairs audio with accurate captions as positives and uses each recording's inaccurate captions and captions from other samples as negatives (Appendix B.2). We use the calibrated similarity $s_{\boldsymbol{a},\boldsymbol{t}} = \alpha \boldsymbol{a}^\intercal \boldsymbol{t} + b$, with trainable scale α and bias b.

Audio-to-text supervision. An audio anchor's accurate captions should receive higher scores than inaccurate ones. For valid audio queries $\mathcal{I}_A = \left\{ i : \left|\mathcal{T}_i^+\right| > 0,\ \left|\mathcal{T}_i^-\right| > 0 \right\}$, the SupCon loss is

$$\mathcal{L}_{\text{SupCon}}^{A\to T} = -\frac{1}{|\mathcal{I}_A|} \sum_{i \in \mathcal{I}_A} \frac{1}{\left|\mathcal{T}_i^+\right|} \sum_{\boldsymbol{t}_i^+ \in \mathcal{T}_i^+} \log \frac{\exp s_{\boldsymbol{a}_i,\boldsymbol{t}_i^+}}{\sum_{\boldsymbol{t}_i \in \mathcal{T}_i^+ \cup \mathcal{T}_i^-} \exp s_{\boldsymbol{a}_i,\boldsymbol{t}_i}}. \tag{2}$$

Text-to-audio supervision. Since an accurate caption's paired recording should receive a higher score than mined negatives, for valid caption queries $\mathcal{I}_T = \left\{ i : \left|\mathcal{T}_i^+\right| > 0,\ |\mathcal{M}_i| > 0 \right\}$, the loss is

$$\mathcal{L}_{\text{SupCon}}^{T\to A} = -\frac{1}{\sum_{i \in \mathcal{I}_T} \left|\mathcal{T}_i^+\right|} \sum_{i \in \mathcal{I}_T} \sum_{\boldsymbol{t}_i^+ \in \mathcal{T}_i^+} \log \frac{\exp s_{\boldsymbol{a}_i,\boldsymbol{t}_i^+}}{\exp s_{\boldsymbol{a}_i,\boldsymbol{t}_i^+} + \sum_{\boldsymbol{a}_{\text{mined}} \in \mathcal{M}_i} \exp s_{\boldsymbol{a}_{\text{mined}},\boldsymbol{t}_i^+}}. \tag{3}$$

Because the query-level averaging assigns equal weight to recordings in the audio-to-text direction and to captions in the text-to-audio direction, recordings with more accurate captions contribute more to Eq. 3. Additionally, mining audio negatives from the same batch for text-to-audio supervision avoids increasing batch size and allows negative selection to adapt as the representations change (Appendix D.1).

Unlike audio-to-text supervision, which supports multiple accurate captions, identifying additional positive recordings is more difficult for text-to-audio supervision. Simple approaches include data augmentation and finding samples with exact caption matches in the training corpus. Preliminary experiments show negligible improvements across tasks when we introduce positive audio samples with these methods. Thus, we exclude additional positive audio from the text-to-audio local supervision in Eq. 3.

Combined objective.

$$\mathcal{L}_{\text{SupCon}} = \frac{1}{2}\left(\mathcal{L}_{\text{SupCon}}^{A\to T} + \mathcal{L}_{\text{SupCon}}^{T\to A}\right), \tag{4}$$

$$\mathcal{L}_{\text{SLS}} = \mathcal{L}_{\text{SigLIP}} + \lambda \mathcal{L}_{\text{SupCon}}, \tag{5}$$

where $\lambda = 1$ by default. After training, the audio-text scoring interface supports retrieval, classification, and ranking through embedding similarities. The same interface scores TTS instruction adherence by comparing synthesized speech with text instructions.

3 EXPERIMENTS

We evaluate SLS on speech tasks across three backbone families (Sections 3.1 and 3.3). Component and mining studies examine the training design (Section 3.4). Finally, qualitative examples and score distributions illustrate the caption discrimination capability (Section 3.5).

3.1 SPEECH-CENTERED EVALUATION

Retrieval and classification. We measure audio-text retrieval with Recall@5 on ParaSpeechCaps (Diwan et al., 2025). Classification selects the highest-scoring text prompt label and reports macro F1 for four speech families: *Speaking Style* (gender, age, accent, pitch, speaking rate, fluency, and emotion), *Recording Conditions* (distance and reverberation), *Conversation* (overlap, turn count, dialog act, and backchannel detection), and *Speech Content* (keywords and intents). Following Vyas et al. (2026), we use Dual Softmax reweighting (Cheng et al., 2021) with scale 4 and a uniform target prior for retrieval and classification, while ranking uses raw similarity.

Ranking. *Caption ranking* evaluates whether a model ranks accurate captions above inaccurate ones with single-attribute contradictions for a recording query. This task resembles SPAM's positive-over-negative adherence metric (Cho et al., 2026). *Audio ranking* evaluates whether a model ranks recordings with matching speaking-style tags above recordings with contrasting tags for a caption query. Each ranking query has five positive and five negative candidates. For query $\boldsymbol{q}$ and nonempty positive and negative evaluation sets $\mathcal{P}(\boldsymbol{q})$ and $\mathcal{N}(\boldsymbol{q})$, we define pairwise accuracy (PairAcc) to measure how often positives outrank negatives:

$$\begin{aligned} \text{PairAcc}(\boldsymbol{q}) &= \frac{1}{|\mathcal{P}(\boldsymbol{q})|\,|\mathcal{N}(\boldsymbol{q})|} \sum_{\boldsymbol{p}\in\mathcal{P}(\boldsymbol{q})} \sum_{\boldsymbol{n}\in\mathcal{N}(\boldsymbol{q})} h(\boldsymbol{q},\boldsymbol{p},\boldsymbol{n}), \\ h(\boldsymbol{q},\boldsymbol{p},\boldsymbol{n}) &= \mathbb{1}[\mathrm{sim}(\boldsymbol{q},\boldsymbol{p}) > \mathrm{sim}(\boldsymbol{q},\boldsymbol{n})] + \frac{1}{2}\mathbb{1}[\mathrm{sim}(\boldsymbol{q},\boldsymbol{p}) = \mathrm{sim}(\boldsymbol{q},\boldsymbol{n})]\,, \end{aligned} \tag{6}$$

where sim and $\mathbb{1}$ denote embedding similarity and the indicator function, respectively. E.g., in caption ranking, $\boldsymbol{q}$ is an audio and $\mathcal{P}(\boldsymbol{q})$ and $\mathcal{N}(\boldsymbol{q})$ are accurate and inaccurate caption sets, respectively. The expectation is 0.5 under random ordering.

Extending this ranking task, Table 12 in Appendix C.4 scores only the original accurate caption or paired recording against the same five negatives, removing the other four positive candidates from each existing ranking. We measure performance with Mean Rank and Top-1 accuracy.

3.2 TRAINING AND COMPARATORS

Training. We combine multiple English speech, sound, and music datasets with LLM-generated accurate and inaccurate captions. We fine-tune VoiceCLAP S, M2D-CLAP, and PE_{AV}S/B/L, with video-related components removed in PE_{AV} models. ESLTTS and CANDOR use separate speakers for training and evaluation, while other datasets follow the respective partitioning protocols (Wang et al., 2024; Reece et al., 2023). DailyTalk uses recording-disjoint train/validation partitions with shared speakers (Lee et al., 2023). The evaluation sets are disjoint from training data.

Comparators. We compare SLS with pretrained audio-text models, including VoiceCLAP (LAION, 2026), ParaSpeechCLAP (Diwan et al., 2026), CLSP (Yang et al., 2026), and AuroLA 7B (Xu et al., 2026). Audio-LLM comparators include Audio Flamingo Next (AF-Next) (Ghosh et al., 2026), Kimi-Audio (KimiTeam et al., 2025), Qwen3-Omni (Xu et al., 2025), and Gemini 3.7 Flash. We omit SPAM from empirical comparisons because the code and pretrained weights were unavailable at the time of evaluation (Cho et al., 2026). Appendices A, B, and C offer experiment details and full results.

3.3 MAIN RESULTS

Retrieval. As shown in Table 1, SLS improves ParaSpeechCaps retrieval for all five pretrained models. All adapted PE_{AV} variants exceed CLSP, with larger variants achieving higher recall. ParaSpeechCLAP has the highest recall averaged across the two directions, while SLS-adapted PE_{AV} models achieve competitive speech retrieval using the same fine-tuning method across model sizes.

Table 1: Comparison of speech-focused retrieval, classification, and ranking. We convert classification problems to multiple-choice QA for audio LLMs. For caption ranking, we feed each audio LLM a query audio and ten captions, labeled A to J, while audio ranking follows the same format. All scores are on a 0 to 100 scale. ▲ indicates absolute changes from each pretrained backbone.

	Model Parameters		Retrieval (Recall@5)			Classification (Macro F1)					Ranking (PairAcc)		
Model	Audio	Text	Audio→Text	Text→Audio	Average	Speaking Style	Recording Conditions	Conversation	Content	Average	Caption	Audio	Average
Random Guess	–	–	0.9	0.9	0.9	23.1	33.3	37.5	3.3	24.3	50.0	50.0	50.0
Audio-text Models													
VoiceCLAP S	0.09B	0.02B	6.4	4.9	5.7	35.8	20.8	34.7	3.4	23.7	70.8	60.1	65.5
+ **SLS**	0.09B	0.02B	55.5	58.4	(▲51.3) 57.0	49.6	64.3	56.2	79.2	(▲38.6) 62.3	96.8	79.4	(▲22.6) 88.1
M2D-CLAP[25']	0.09B	0.1B	3.6	3.3	3.5	33.1	45.9	35.4	3.1	29.4	56.4	55.8	56.1
+ **SLS**	0.09B	0.1B	61.3	62.9	(▲58.6) 62.1	45.2	67.4	61.3	86.3	(▲35.6) 65.0	96.5	80.5	(▲32.4) 88.5
PE_{AV}S	0.09B	0.3B	15.5	16.0	15.8	46.8	51.5	44.5	46.5	47.3	65.1	64.8	65.0
+ **SLS**	0.09B	0.3B	73.6	74.0	(▲58.0) 73.8	**52.7**	**86.2**	65.1	91.9	(▲26.7) 74.0	97.3	**81.5**	(▲24.4) **89.4**
PE_{AV}B	0.2B	0.3B	16.9	15.8	16.4	46.3	48.1	38.8	49.3	45.6	67.0	62.0	64.5
+ **SLS**	0.2B	0.3B	77.5	77.8	(▲61.3) 77.7	52.3	82.9	64.5	94.6	(▲28.0) 73.6	97.3	79.1	(▲23.7) 88.2
PE_{AV}L	1B	0.3B	15.6	17.1	16.4	48.2	58.9	41.8	50.3	49.8	61.5	65.2	63.4
+ **SLS**	1B	0.3B	78.0	**78.4**	(▲61.8) 78.2	50.8	85.5	**68.2**	96.2	(▲25.4) **75.2**	97.7	77.7	(▲24.3) 87.7
ParaSpeechCLAP	0.3B	0.3B	**81.1**	**78.4**	**79.8**	35.3	33.0	32.9	2.1	25.8	63.4	**81.5**	72.5
CLSP	0.6B	0.1B	68.9	67.5	68.2	43.1	33.4	41.3	5.8	30.9	64.7	**81.5**	73.2
VoiceCLAP L	0.6B	7B	15.1	12.7	13.9	45.6	41.1	51.6	79.7	54.5	70.4	65.0	67.7
AuroLA 7B	0.6B	7B	12.9	11.3	12.1	42.7	38.9	54.3	95.4	57.8	69.1	67.7	68.4
Audio LLMs													
AF-Next	0.6B	7B	–	–	–	30.0	25.8	45.9	86.3	47.0	69.5	52.4	60.9
Kimi-Audio	0.6B	7B	–	–	–	34.5	35.7	39.0	96.0	51.3	79.6	60.3	69.9
Qwen3-Omni	0.6B	30B	–	–	–	38.3	27.1	62.6	97.1	56.3	85.1	61.7	73.4
Gemini 3.7 Flash	–	–	–	–	–	42.4	53.4	66.6	**98.2**	65.1	**99.8**	77.9	88.9

Classification. All five SLS-adapted models improve the average macro F1 in each of the four speech classification families over pretrained models, exceeding CLSP and ParaSpeechCLAP in every family. Adapted PE_{AV} models achieve the highest overall classification scores among the listed comparators. Moreover, task-level results in Table 10 show improvements in accent, fluency, recording distance, and backchannel detection, implying the usefulness of the caption supervision for recognizing vocal attributes, recording conditions, and conversational behavior.

Ranking. SLS improves caption and audio ranking over all five corresponding pretrained models. All five adaptations exceed CLSP and ParaSpeechCLAP in caption ranking PairAcc. Additionally, Appendix C.6 shows that the retrieval, classification, and ten-candidate ranking gains persist across three fine-tuning seeds for VoiceCLAP S, M2D-CLAP, and PE_{AV}S. Table 12 further evaluates ranking of the original caption or paired recording against five closely related negatives with contrasting attributes. All five SLS adaptations lower Mean Rank and increase Acc@1 in both ParaSpeechCaps audio-text directions relative to pretrained models. Caption Acc@1 reaches 71.8% to 77.2%, compared with 42.1% for ParaSpeechCLAP, the strongest unadapted audio-text model on this measure. Therefore, these gains extend to identifying the original accurate caption among inaccurate captions with targeted attribute contradictions.

Following the paired evaluation protocol of Ghosh et al. (2024), Appendix C.5 examines whether models correctly match two recordings from the same speaker to their respective captions describing contrasting speaking styles. Table 13 reports results on manually reviewed ParaSpeechCaps pairs. Compared with the pretrained models, all SLS adaptations more often correctly match both recordings and both captions for speaking rate, emotion, and animated versus calm/sleepy speech.

Beyond English speech, Coco-Nut evaluates Japanese speech using English consensus captions derived from Japanese listener descriptions. As shown in Table 11, SLS improves PairAcc for four of

Table 2: Component study across three pretrained backbones. Retrieval (Rtr) averages Recall@5 over both audio-text directions. Classification (Cls) averages macro F1 over the four speech families. Ranking (Rnk) averages PairAcc over both audio-text directions. Rows up to (4) add caption supervision cumulatively. Rows (5) to (7) add different SupCon setups to row (4) respectively.

	VoiceCLAP S			M2D-CLAP			PE_{AV}S		
Method	Rtr	Cls	Rnk	Rtr	Cls	Rnk	Rtr	Cls	Rnk
(1) Pretrained Checkpoint	5.7	23.7	65.5	3.5	29.4	56.1	15.8	47.3	65.0
(2) + Standard SigLIP Fine-tuning	**70.3**	57.1	73.0	**75.1**	63.2	73.3	**85.0**	71.4	73.9
(3) + Accurate Captions	61.5	59.5	79.3	66.4	63.6	78.1	78.6	72.2	80.9
(4) + Inaccurate Captions	57.4	61.0	87.9	57.7	60.5	87.8	75.7	71.8	88.7
(5) (4) + Audio→Text SupCon	17.8	59.4	79.4	11.8	63.3	78.8	25.2	71.5	82.1
(6) (4) + Text→Audio SupCon	58.4	58.9	85.9	59.2	63.4	88.3	73.3	70.7	88.1
(7) (4) + Bidirectional SupCon (**SLS**)	57.0	**62.3**	**88.1**	62.1	**65.0**	**88.5**	73.8	**74.0**	**89.4**

Table 3: Effect of in-batch audio-negative selection on VoiceCLAP S under the full SLS objective. All scores are on a 0 to 100 scale. Similarity-based strategies use mean similarity to the anchor recording's inaccurate captions, while random selection samples uniformly without replacement. Median similarity selects a contiguous group centered in the similarity-sorted candidate list.

Audio-negative Selection	Retrieval	Classification	Caption Ranking	Audio Ranking
Uniform Random	42.1	60.7	97.5	76.8
Lowest Similarity	14.1	58.2	**98.5**	63.2
Median Similarity	17.2	59.7	98.3	66.6
Highest Similarity (**SLS**)	**57.0**	**62.3**	96.8	**79.4**

five models and StrictSep for all five, where StrictSep measures the percentage of queries for which every accurate caption ranks above every inaccurate one. These results complement ParaSpeechCaps by extending the evidence for local speech discrimination to a different spoken language and caption source. Together, these evaluation results demonstrate SLS's effectiveness in distinguishing accurate captions and recordings from closely related inaccurate alternatives.

3.4 Component and Mining Studies

Caption relations and adaptation. Table 2 compares standard SigLIP fine-tuning (row 2), adding accurate and inaccurate captions (rows 3 and 4), and adding SupCon losses (rows 5 to 7). Standard SigLIP achieves the strongest retrieval, while accurate and inaccurate captions progressively improve ranking across all three backbones, with some trade-off in retrieval performance.

Bidirectional local supervision. With the same accurate and inaccurate captions, SLS improves average classification and ranking over SigLIP alone and both single-direction configurations across all three backbones (rows 7 vs. 4 to 6). Note that rows 5 and 6 retain the $1/2$ coefficient from Eq. 4. Notably, retrieval drops sharply with audio-to-text SupCon alone (row 5) but is stronger when text-to-audio supervision is included (row 7), showing the importance of the proposed bidirectional local supervision with in-batch audio negatives. Therefore, the component study demonstrates the efficacy of the proposed structured caption relations and local discrimination supervision.

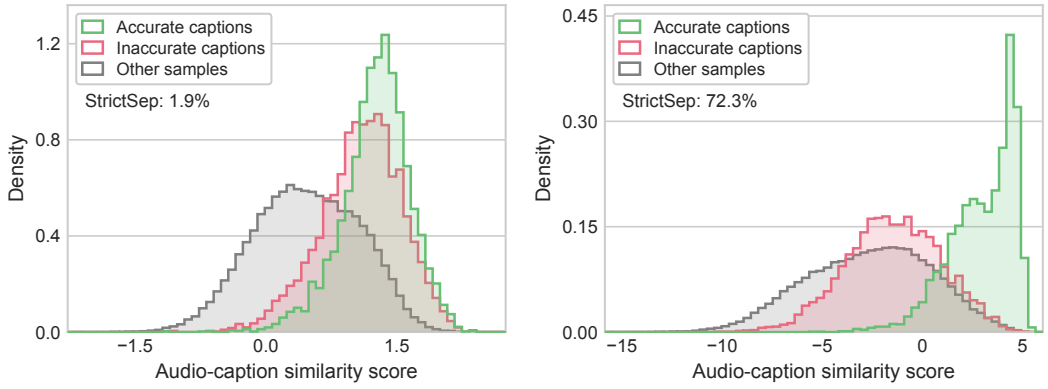
In-batch audio-negative mining. Table 3 compares audio-negative selection strategies for VoiceCLAP S under the full SLS objective. Selecting recordings with the highest average similarity to the current recording's inaccurate captions offers the strongest retrieval, classification, and audio ranking scores. Furthermore, retrieval and classification improve as the selected audio-negatives move towards the top of the similarity-sorted candidate list (lowest to median to highest), indicating that the similarity metric in Eq. 1 is meaningful and successfully retrieves useful audio negatives from the training batch. We report additional selection methods in Appendix D.1.

Table 4: Successful ParaSpeechCaps caption ranking example. **Bold red** ranks indicate inaccurate captions that outrank at least one accurate caption. Full results in Table 18.

Candidate	Rank $PE_{AV}S$	Rank **SLS**
Ground Truth	3	5
Paraphrase	5	1
Reordering	8	4
Label-style	6	3
Partial Positive	10	2
Inaccurate (medium → high-pitch)	**7**	6
Inaccurate (noisy → clean)	**1**	7
Inaccurate (measured → rapid)	**2**	8
Inaccurate (silky → rough)	**4**	9
Inaccurate (female → male)	**9**	10

Table 5: Failed ParaSpeechCaps caption ranking example. **Bold red** ranks indicate inaccurate captions that outrank at least one accurate caption. Full results in Table 19.

Candidate	Rank $PE_{AV}S$	Rank **SLS**
Ground Truth	7	7
Paraphrase	6	3
Reordering	9	5
Label-style	3	4
Partial Positive	1	2
Inaccurate (British → American)	**5**	**1**
Inaccurate (high → low-pitch)	**2**	**6**
Inaccurate (very noisy → clear)	**8**	8
Inaccurate (measured → rapid)	**4**	9
Inaccurate (female → male)	10	10



(a) $PE_{AV}S$ (b) SLS-$PE_{AV}S$

Figure 2: Density distributions of audio-text dot-product scores ($\boldsymbol{a}^{\top}\boldsymbol{t}$) on ParaSpeechCaps for (a) pretrained $PE_{AV}S$ and (b) SLS-adapted $PE_{AV}S$. Accurate captions include original reference descriptions, reordered and paraphrased variants, label-style captions, and partial but correct descriptions. "Other samples" pools captions associated with unrelated recordings. Each distribution is normalized to unit area. (a) and (b) use different score ranges.

3.5 Qualitative Examples and Score Analysis

Qualitative caption ranking. Tables 4 and 5 illustrate caption ranking before and after SLS adaptation of $PE_{AV}S$. In Table 4, the pretrained $PE_{AV}S$ ranks clean-environment and rapid-speech contradictions first and second, while the valid partial description ranks last. Conversely, SLS places all five accurate descriptions above every inaccurate one, accepting variations in wording, format, and detail while distinguishing attribute contradictions. Next, Table 5 illustrates remaining errors after SLS adaptation. SLS lowers the low-pitch and rapid-speech inaccurate captions but ranks the American accent contradiction first. Both the accent and low-pitch contradictions outrank the original caption, showing that some attribute contrasts remain difficult after adaptation.

Similarity score distribution. Figure 2 compares audio-caption similarity distributions before and after SLS adaptation. SLS reduces overlap between accurate and inaccurate caption scores. StrictSep increases from 1.9% to 72.3%, indicating that all five accurate captions outrank all five inaccurate captions for most recordings. Furthermore, inaccurate captions generally score higher than captions from other recordings, suggesting that targeted attribute contrasts offer harder alternatives. Overall, the reduced overlap and higher StrictSep support stronger local discrimination.

Table 6: Spearman correlations between automatic adherence scores and mean listener ratings for each instruction type. Macro Avg averages APS, DSD, and RP correlations equally. Displayed ± margins give the maximum distance from the estimate to either bound of the 95% utterance-clustered bootstrap interval. ▲ denotes macro gains over the corresponding pretrained model, and ** denotes Holm-adjusted $p < 0.01$. **Bold** indicates the highest correlation in each column and evaluator type.

Automatic Judgment	APS	DSD	RP	Macro Avg
Audio-text Models				
M2D-CLAP	0.163±0.195	0.025±0.191	0.118±0.212	0.102±0.136
+ SigLIP with Structured Captions	0.482±0.156	**0.346**±0.128	0.285±0.190	(▲0.269**) 0.371±0.101
+ **SLS**	0.371±0.146	0.328±0.129	0.340±0.148	(▲0.245**) 0.347±0.086
VoiceCLAP S	**0.578**±0.120	0.086±0.187	**0.397**±0.128	0.354±0.105
+ SigLIP with Structured Captions	0.551±0.110	0.320±0.149	0.230±0.193	(▲0.013) 0.367±0.105
+ **SLS**	0.565±0.118	0.336±0.127	0.309±0.192	(▲0.049) **0.403**±0.092
$PE_{AV}S$	0.551±0.132	0.022±0.149	−0.017±0.167	0.185±0.100
+ SigLIP with Structured Captions	0.477±0.145	0.193±0.162	0.355±0.119	(▲0.157) 0.342±0.102
+ **SLS**	0.560±0.100	0.271±0.129	0.375±0.157	(▲0.217**) **0.402**±0.082
ParaSpeechCLAP	0.390±0.134	0.192±0.173	0.246±0.186	0.276±0.100
CLSP	0.552±0.139	0.199±0.209	0.257±0.162	0.336±0.117
Audio LLMs				
AF-Next	0.136±0.107	0.124±0.111	0.152±0.129	0.137±0.063
Kimi-Audio	0.136±0.135	0.171±0.167	0.313±0.195	0.207±0.108
Qwen3-Omni	0.287±0.176	0.033±0.046	0.157±0.224	0.159±0.102
Gemini 3.7 Flash	0.467±0.145	0.379±0.170	0.360±0.151	0.402±0.100
Gemini 3.1 Pro Preview	**0.502**±0.122	**0.383**±0.173	**0.416**±0.151	**0.434**±0.094

4 Transfer to TTS Instruction-Adherence Evaluation

4.1 Setup

We evaluate how well automatic scores agree with listeners on whether speech follows a given instruction (Lin et al., 2026). Building on InstructTTSEval's validation of automatic judges using reference audio, we collect listener ratings for both synthesized and reference speech (Huang et al., 2025). We consider three instruction types: Acoustic-Parameter Specification (APS) lists 12 explicit acoustic attributes, Descriptive-Style Directive (DSD) describes the desired speaking style in natural language, and Role-Play (RP) specifies a role or scenario.

We select 20 source utterances from the English subset of InstructTTSEval. For each source utterance and instruction type, we evaluate the reference recording and five TTS outputs: Gemini 3.1 Flash TTS Preview with and without instructions, Hume Octave, Qwen3-TTS VoiceDesign 1.7B (Hu et al., 2026), and MOSS-VoiceGenerator (Huang et al., 2026). The 20 utterances, six audio conditions, and three instruction types yield 360 evaluation examples. We collect 1,800 ratings from 150 listeners, with 3 to 6 ratings per example. For APS, listeners mark which of 12 attributes match the recording, and the fraction of matching attributes gives a score between 0 and 1. For DSD and RP, listeners rate how well the speech follows the instruction on a scale from 1 to 5. Appendix F describes the synthesis and listening procedures, and Table 20 reports the listener ratings.

Audio-text model evaluators compute the raw similarity between audio and full instruction embeddings. Audio-LLM judges use the same scoring scales as human listeners. We correlate automatic scores with listener means using Spearman correlation over 120 utterance-condition combinations per instruction type. Macro correlation averages the three instruction type correlations equally. No listener ratings or evaluation examples were used during SLS adaptation.

4.2 Results

Transfer from pretrained models. In Table 6, SLS improves correlation with listener adherence ratings, with significant gains for M2D-CLAP and $PE_{AV}S$. Compared with SigLIP using the same

structured captions, SLS achieves higher average correlation for VoiceCLAP S and $PE_{AV}S$. Results support using the adapted models to judge how well speech follows the requested delivery or style.

Comparison with other automatic judges. All three SLS evaluators exceed CLSP and ParaSpeechCLAP in observed DSD, RP, and macro correlations. Adapted VoiceCLAP S leads audio-text evaluators at 0.403 macro correlation, followed by SLS-adapted $PE_{AV}S$ at 0.402, while Gemini 3.1 Pro Preview leads overall at 0.434. Despite having up to 100× fewer parameters than audio LLMs like Qwen3-Omni (30B) or AF-Next (7B), SLS-adapted models (0.1B-0.4B) achieve substantially higher correlation with human perceptual ratings across DSD and RP instructions. While LLMs provide qualitative feedback, audio-text models with dual encoders offer reusable embeddings and efficient batch scoring over thousands of synthesis outputs. Pearson correlations in Table 21 show the same broad pattern.

5 Related Work

Audio-text representations. Contrastive audio-text models support retrieval and zero-shot classification (Wu et al., 2023; Elizalde et al., 2023; Niizumi et al., 2025). ParaCLAP and ParaSpeechCLAP extend to paralinguistic attributes like speaking style descriptions (Jing et al., 2024; Diwan et al., 2026). Auden-Voice studies retrieval and voice understanding through multitask voice supervision (Huo et al., 2026). CLSP combines global and fine-grained positive captions for speech-text representation learning (Yang et al., 2026).

Fine-grained alignment. SPAM combines style-defined positives, bidirectional SupCon, and prompt-adherence evaluation (Cho et al., 2026). RA-CLAP learns self-distilled soft relations for emotional speaking-style retrieval, while EmotionRankCLAP uses valence-arousal orderings (Sun et al., 2025; Chandra et al., 2025). Counterfactual audio learning uses caption edits (Vosoughi et al., 2024), and CompA and ARO/NegCLIP use composition-aware negatives to strengthen discrimination beyond retrieval (Ghosh et al., 2024; Yuksekgonul et al., 2023). AuroLA combines multi-granular captions with tag-defined positives and hard-negative weighting (Xu et al., 2026). SLS trains existing audio-text models with accurate and inaccurate descriptions, where the same inaccurate captions guide selection of audio negatives for text-to-audio SupCon.

Instruction-conditioned TTS evaluation. Instruction-following benchmarks and perceptual studies evaluate TTS (Huang et al., 2025; Chen et al., 2026b; Lin et al., 2026). Learned audio-text scores include SPAM's style-adherence metric, ParaSpeechCLAP's candidate scoring, and models trained on human relevance judgments (Cho et al., 2026; Diwan et al., 2026; Takano et al., 2025). We extend InstructTTSEval's reference audio judge validation with listener ratings of synthesized and reference speech, testing SLS's transfer to automatic instruction-adherence scoring.

6 Conclusions

SLS trains audio-text models with constructed captions, uses inaccurate captions to select audio negatives, and applies SupCon in both audio-to-text and text-to-audio directions. Across three model families, SLS improves speech retrieval, classification, and ranking. Component and qualitative studies support the caption supervision and the additional losses. The listening study shows improved average correlation with human judgments of TTS instruction adherence, extending the application beyond retrieval and classification.

Limitations and future work. Although Coco-Nut offers evidence of Japanese speech transfer, broader multilingual evaluation and non-English text alignment remain unexplored. Our compute budget limits exploration of larger training scales and pretraining from scratch. Moreover, SLS defines locality through caption relations and uses recording-level similarity scores, while frame-level understanding and temporal localization remain unexplored. Finally, while this work demonstrates adherence evaluation in a post-hoc setting, the continuous and differentiable nature of SLS audio-text representations makes them natural candidates for reinforcement learning with AI feedback or direct preference optimization in expressive TTS without invoking multi-billion-parameter audio LLMs at training time (Chen et al., 2026a).

AI USE STATEMENT

We used generative AI tools for data construction (pseudo-labeling, synthetic caption generation, translation), coding, literature review, and manuscript editing. We reviewed all AI-assisted work and verified and tested generated code. We take responsibility for the final text, claims, and artifacts.

ETHICS STATEMENT

The evaluation uses source speech datasets and human listener judgments. Appendix A describes annotation provenance, and Appendix F describes the listening protocol. SLS may improve speech retrieval and the evaluation of expressive speech synthesis. However, descriptions of perceived speaker characteristics, particularly personality, emotion, age, gender, and accent, may encode annotator biases and should not be treated as verified facts about speakers. Broader deployment requires evaluating performance across languages and speaker groups. These representations could also facilitate unwanted speaker profiling, while improvements in expressive synthesis may be misused to produce deceptive audio. Finally, using SLS scores as optimization targets requires care, as systems may exploit evaluator weaknesses. Hence, automatic scores should complement human assessment.

REPRODUCIBILITY STATEMENT

Section 2 defines SLS's construction and objectives. Appendices A and B describe evaluation tasks and training settings, and Appendix F.2 specifies the reported uncertainty and pretrained-comparison procedures. We plan to make the SLS-adapted audio-text models and the LLM-generated structured captions for ParaSpeechCaps publicly available.

REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *LREC*, 2020.

Carlos Busso, Reza Lotfian, Kusha Sridhar, Ali N. Salman, Wei-Cheng Lin, Lucas Goncalves, Srinivas Parthasarathy, Abinay Reddy Naini, Seong-Gyun Leem, Luz Martinez-Lucas, Huang-Cheng Chou, and Pravin Mote. The msp-podcast corpus. *IEEE Transactions on Affective Computing*, 17 (3), 2026.

H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4), 2014.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain A. McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The AMI meeting corpus: A pre-announcement. In *MLMI*, 2005.

Shreeram Suresh Chandra, Lucas Goncalves, Junchen Lu, Carlos Busso, and Berrak Sisman. Emotionrankclap: Bridging natural language speaking styles and ordinal speech emotion via rank-n-contrast. In *Interspeech*, 2025.

Dekun Chen, Xueyao Zhang, Yuancheng Wang, Kenan Dai, Li Ma, and Zhizheng Wu. Flexivoice: Enabling flexible style control in zero-shot tts with natural language instructions. In *ICLR*, 2026a.

Huakang Chen, Jingbin Hu, Liumeng Xue, Qirui Zhan, Wenhao Li, Guobin Ma, Hanke Xie, Dake Guo, Linhan Ma, Yuepeng Jiang, et al. Mint-bench: A comprehensive multilingual benchmark for instruction-following text-to-speech. *arXiv preprint arXiv:2604.17958*, 2026b.

Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021.

Chanhee Cho, Nayeon Kim, and Bugeun Kim. Spam: Style prompt adherence metric for prompt-based tts. In *ICASSP*, 2026.

Heinrich Dinkel, Zhiyong Yan, Tianzi Wang, Yongqing Wang, Xingwei Sun, Yadong Niu, Jizhong Liu, Gang Li, Junbo Zhang, and Jian Luan. Glap: General contrastive audio-text pretraining across domains and languages. In *ICASSP*, 2026.

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

Anuj Diwan, Eunsol Choi, and David Harwath. Paraspeechclap: A dual-encoder speech-text model for rich stylistic language-audio pretraining. In *Interspeech*, 2026.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP*, 2020.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: learning audio concepts from natural language supervision. In *ICASSP*, 2023.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. Natural language supervision for general-purpose audio representations. In *ICASSP*, 2024.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with WaveNet autoencoders. In *ICML*, 2017.

Tiantian Feng, Jihwan Lee, Anfeng Xu, Yoonjeong Lee, Thanathai Lertpetchpun, Xuan Shi, Helin Wang, Thomas Thebaud, Laureano Moro-Velazquez, Dani Byrd, et al. Vox-profile: A speech foundation model benchmark for characterizing diverse speaker and speech traits. *arXiv preprint arXiv:2505.14648*, 2025.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Reddy Evuru, Ramaneswaran S, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Compa: Addressing the gap in compositional reasoning in audio-language models. In *ICLR*, 2024.

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

Yuan Gong, Jin Yu, and James Glass. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP*, 2022.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

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

Kexin Huang, Liwei Fan, Botian Jiang, Yaozhou Jiang, Qian Tu, Jie Zhu, Yuqian Zhang, Yiwei Zhao, Chenchen Yang, Zhaoye Fei, et al. Moss-voicegenerator: Create realistic voices with natural language descriptions. *arXiv preprint arXiv:2603.28086*, 2026.

Mingyue Huo, Wei-Cheng Tseng, Yiwen Shao, Hao Zhang, and Dong Yu. Auden-voice: General-purpose voice encoder for speech and language understanding. In *ICASSP*, 2026.

Jisu Jeon, Seungyeon Jwa, Joosung Lee, Jinhyeon Kim, Woojin Chung, Hwiyeol Jo, Jeonghoon Kim, Jonghyun Choi, and Soyoon Kim. Parapairaudiobench: Paralinguistic pairwise audio benchmark for lalm-as-a-judge. *arXiv preprint arXiv:2606.24648*, 2026.

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

Amita Kamath, Cheng-Yu Hsieh, Kai-Wei Chang, and Ranjay Krishna. The hard positive truth about vision-language compositionality. In *ECCV*, 2024.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *NeurIPS*, 2020.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *NAACL*, 2019.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report, 2025.

LAION. Laionbox: Fine-tuning dramabox tts with multi-auxiliary differentiable losses, 2026. URL `https://github.com/LAION-AI/laionbox`.

Keon Lee, Kyumin Park, and Daeyoung Kim. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP*, 2023.

Guojian Li, Zhixian Zhao, Zhennan Lin, Jingbin Hu, Qirui Zhan, Yuang Cao, Pengyuan Xie, Chuan Xie, Jie Liu, Qiang Zhang, et al. Towards fine-grained multi-dimensional speech understanding: Data pipeline, benchmark, and model. *arXiv preprint arXiv:2605.12036*, 2026a.

Haitao Li, Tian Tan, Yuguang Yang, Shan Yang, and Xie Chen. Anyaudio-judge: A dynamic rubric-based benchmark and evaluator for audio instruction following. *arXiv preprint arXiv:2606.03116*, 2026b.

Xiquan Li, Xuenan Xu, Ziyang Ma, Wenxi Chen, Haolin He, Qiuqiang Kong, and Xie Chen. FineLAP: Taming heterogeneous supervision for fine-grained language-audio pretraining. In *ACL*, 2026c.

Yiming Li, Zhifang Guo, Xiangdong Wang, and Hong Liu. Advancing multi-grained alignment for contrastive language-audio pre-training. In *ACM MM*, 2024.

Yi-Cheng Lin, Huang-Cheng Chou, Tzu-Chieh Wei, Kuan-Yu Chen, and Hung-yi Lee. Do you hear what i mean? quantifying the instruction-perception gap in instruction-guided expressive text-to-speech systems. In *ICASSP*, 2026.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. Speech model pre-training for end-to-end spoken language understanding. In *Interspeech*, 2019.

Max Morrison, Caedon Hsieh, Nathan Pruyne, and Bryan Pardo. Cross-domain neural pitch and periodicity estimation. *arXiv preprint arXiv:2301.12258*, 2023.

Daisuke Niizumi, Daiki Takeuchi, Masahiro Yasuda, Binh Thien Nguyen, Yasunori Ohishi, and Noboru Harada. M2d-clap: Exploring general-purpose audio-language representations beyond clap. *IEEE Access*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *ACM MM*, 2015.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science advances*, 9(13), 2023.

Haoqin Sun, Jingguang Tian, Jiaming Zhou, Hui Wang, Jiabei He, Shiwan Zhao, Xiangyu Kong, Desheng Hu, Xinkang Xu, Xinhui Hu, and Yong Qin. Ra-clap: Relation-augmented emotional speaking style contrastive language-audio pretraining for speech retrieval. In *Interspeech*, 2025.

Taisei Takano, Yuki Okamoto, Yusuke Kanamori, Yuki Saito, Ryotaro Nagase, and Hiroshi Saruwatari. Human-clap: Human-perception-based contrastive language-audio pretraining. In *APSIPA ASC*, 2025.

George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), 2002.

Ali Vosoughi, Luca Bondi, Ho-Hsiang Wu, and Chenliang Xu. Learning audio concepts from counterfactual natural language. In *ICASSP*, 2024.

Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matt Le, Piotr Dollár, Christoph Feichtenhofer, et al. Pushing the frontier of audiovisual perception with large-scale multimodal correspondence learning. In *CVPR*, 2026.

Hui Wang, Jinghua Zhao, Yifan Yang, Shujie Liu, Junyang Chen, Yanzhe Zhang, Shiwan Zhao, Jinyu Li, Jiaming Zhou, Haoqin Sun, Yan Lu, and Yong Qin. SpeechLLM-as-judges: Towards general and interpretable speech quality evaluation. In *ACL*, 2026.

Wenbin Wang, Yang Song, and Sanjay Jha. Usat: A universal speaker-adaptive text-to-speech approach. *TASLP*, 32, 2024.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

Aya Watanabe, Shinnosuke Takamichi, Yuki Saito, Wataru Nakata, Detai Xin, and Hiroshi Saruwatari. Coco-nut: Corpus of japanese utterance and voice characteristics description for prompt-based control. In *ASRU*, 2023.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023.

Chenghao Xiao, Hou Pong Ken Chan, Hao Zhang, Weiwen Xu, Mahani Aljunied, and Yu Rong. Scaling language-centric omnimodal representation learning. *NeurIPS*, 2025.

Huang Xie, Okko Räsänen, and Tuomas Virtanen. On negative sampling for contrastive audio-text retrieval. In *ICASSP*, 2023.

Jilan Xu, Carl Thomé, Danijela Horak, Weidi Xie, and Andrew Zisserman. Scaling audio-text retrieval with multimodal large language models. *arXiv preprint arXiv:2602.18010*, 2026.

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

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92), 2019.

Yifan Yang, Bing Han, Hui Wang, Wei Wang, Ziyang Ma, Long Zhou, Zengrui Jin, Guanrou Yang, Tianrui Wang, Xu Tan, and Xie Chen. Towards fine-grained and multi-granular contrastive language-speech pre-training. In *ACL*, 2026.

Hua Ye, Hang Ding, Siyuan Chen, Yiyang Jiang, Changyuan Zhang, and Xuan Zhang. Aligning by misaligning: Boundary-aware curriculum learning for multimodal alignment. *NeurIPS*, 2025.

HaeJun Yoo, Yongseop Shin, Insung Lee, Myoung-Wan Koo, and Du-Seong Chang. Omni-embed-audio: Leveraging multimodal LLMs for robust audio-text retrieval. In *ACL*, 2026.

Yi Yuan, Zhuo Chen, Xubo Liu, Haohe Liu, Xuenan Xu, Dongya Jia, Yuanzhe Chen, Mark D Plumbley, and Wenwu Wang. T-clap: Temporal-enhanced contrastive language-audio pretraining. In *MLSP*, 2024.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

Jun Zhan, Mingyang Han, Yuxuan Xie, Chen Wang, Dong Zhang, Kexin Huang, Haoxiang Shi, DongXiao Wang, Tengtao Song, Qinyuan Cheng, et al. Vstyle: A benchmark for voice style adaptation with spoken instructions. In *ICASSP*, 2026.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *Interspeech*, 2021.

Leying Zhang, Bowen Shi, Haibin Wu, Bach Viet Do, and Yanmin Qian. Jastin: Aligning llms for zero-shot audio and speech evaluation via natural language instructions. *arXiv preprint arXiv:2605.04505*, 2026.

Yiming Zhang, Zhuokai Zhao, Zhaorun Chen, Zhili Feng, Zenghui Ding, and Yining Sun. Rankclip: Ranking-consistent language-image pretraining. In *ICCV*, 2025a.

Zixing Zhang, Yimeng Wu, Zhongren Dong, Wulong Xiang, Shengfan Shen, and Björn W Schuller. Sse: A speaking style extractor based on fine-grained contrastive learning between speech and descriptive text. In *ICASSP*, 2025b.

A Evaluation Tasks

Table 7 records evaluation sample counts, dataset splits, and annotation methods.

A.1 Retrieval Tasks

Retrieval measures whether audio and text identify the same recording in a full candidate pool. We evaluate speech using ParaSpeechCaps (Diwan et al., 2025), sound using AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020), and music using MusicCaps (Agostinelli et al., 2023). We create a subset from the combined evaluation partition of ParaSpeechCaps by removing duplicated utterances with exact speaking style tag matches.[1] For datasets with multiple reference captions, captions associated with the same recording are valid matches. Audio-to-text retrieval ranks all candidate captions for each recording, and text-to-audio retrieval ranks all candidate recordings for each caption. Recall@5 measures the fraction of queries with a valid match among the top five candidates and serves as the summary metric. Additionally, Table 9 reports Recall@1 and Recall@10 to characterize retrieval at narrower and broader cutoffs.

A.2 Classification Tasks

Classification evaluates the four speech families in Sec. 3.1. Speaking Style covers gender and age with Common Voice (Ardila et al., 2020), accent with ESLTTS (Wang et al., 2024), pitch and speaking rate with VCTK (Yamagishi et al., 2019), fluency with SpeechOcean762 (Zhang et al., 2021), and emotion with CREMA-D (Cao et al., 2014) and MSP-Podcast (Busso et al., 2026). Recording Conditions use controlled distance and reverberation transformations of VCTK recordings. Conversation includes overlap and turn counting from AMI (Carletta et al., 2005), dialog acts from DailyTalk (Lee et al., 2023), and backchannel detection from CANDOR (Reece et al., 2023). Speech Content includes keyword detection with Speech Commands (Warden, 2018) and intent classification with Fluent Speech Commands (Lugosch et al., 2019). Following Diwan et al. (2025), pitch labels use PENN estimates (Morrison et al., 2023). Common Voice age labels use contributor-provided age metadata. Details are shown in Table 7.

The CANDOR evaluation includes 135 manually verified short listener responses (backchannels) and 135 examples without a listener response. For ESLTTS, we construct the training split from the official dev set and the evaluation split from the combined ada and test sets, ensuring speaker disjointness between training and evaluation.[2] Complementary sound tasks use ESC-50 (Piczak, 2015) for environmental sounds and VocalSound (Gong et al., 2022) for non-speech vocal events. Music tasks use NSynth (Engel et al., 2017) for instrument families and GTZAN (Tzanetakis & Cook, 2002) for genres. Audio-text models select the highest-scoring label description, while audio LLMs select from the same label inventory. We report macro F1 for each task. We manually review samples from each category in each dataset to check the correctness of the constructed task data.

A.3 Ranking Tasks

Definition. Let $\boldsymbol{q} \in \mathbb{R}^D$ be a query embedding with nonempty, disjoint positive and negative evaluation sets $\mathcal{P}(\boldsymbol{q}), \mathcal{N}(\boldsymbol{q}) \subset \mathbb{R}^D$. Caption ranking uses an audio query $\boldsymbol{q} = \boldsymbol{a}$ with accurate and inaccurate caption embeddings as candidates, while audio ranking uses a caption query $\boldsymbol{q} = \boldsymbol{t}$ with matching and contrasting audio embeddings. A perfect ordering satisfies

$$\operatorname{sim}(\boldsymbol{q}, \boldsymbol{p}) > \operatorname{sim}(\boldsymbol{q}, \boldsymbol{n}), \quad \forall \boldsymbol{p} \in \mathcal{P}(\boldsymbol{q}),\ \boldsymbol{n} \in \mathcal{N}(\boldsymbol{q}). \tag{7}$$

This ordering gives PairAcc $= 1$ in Eq. 6. Each query has five positive and five negative candidates, constructed separately from the training pools.

Caption Ranking. Caption ranking tests whether accurate descriptions outrank targeted inaccurate captions for a given recording. The candidates comprise the original caption, four accurate variants, and five inaccurate captions. Accurate variants include reordered descriptions, paraphrases, label-style captions, and partial descriptions that retain correct information. We construct tasks from

[1] https://hf.co/datasets/ajd12342/paraspeechclap-eval-combined

[2] https://hf.co/datasets/MushanW/ESLTTS

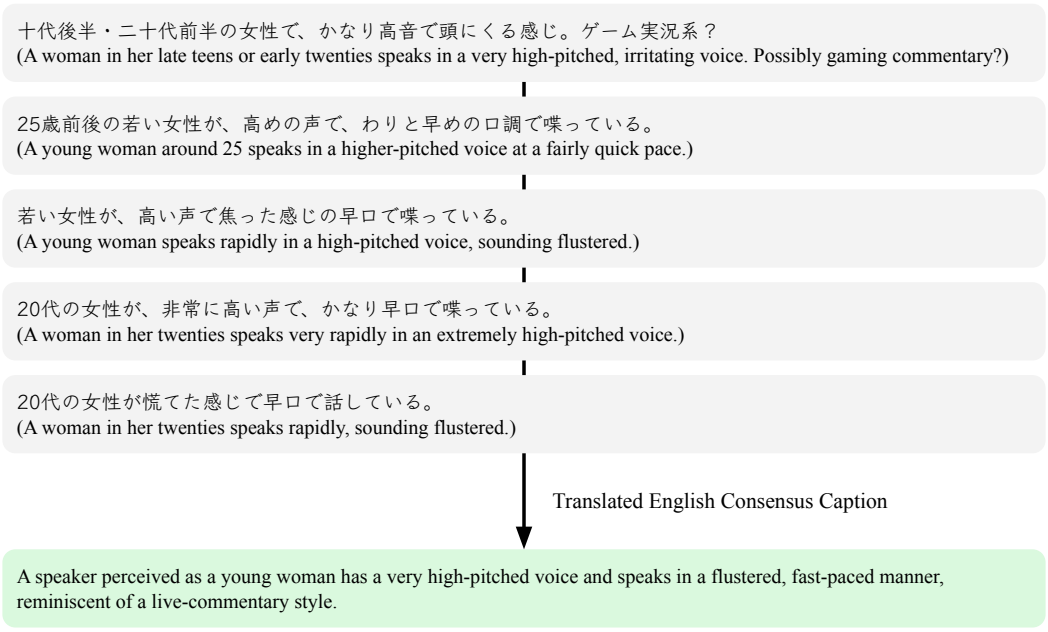


Figure 3: An example of the Japanese listener descriptions in Coco-Nut (Watanabe et al., 2023) and the corresponding LLM-translated English consensus caption using the prompt in Figure 5.

ParaSpeechCaps (Diwan et al., 2025), Coco-Nut (Watanabe et al., 2023), AudioCaps (Kim et al., 2019), and MusicCaps (Agostinelli et al., 2023). Coco-Nut uses Japanese speech with English consensus captions derived from Japanese listener descriptions, extending evaluation beyond the English training setting.

Audio Ranking. Audio ranking tests whether recordings with matching speaking-style outrank recordings with contrasting attributes for a caption. We use the original caption as the query when defining positive audio. Positives comprise the source recording and four recordings with exactly matching style tags. For source tag set $\mathcal{S}$ and candidate tag set $\mathcal{C}$, eligible negatives come from different speakers than the source, satisfy $\mathcal{S} \not\subseteq \mathcal{C}$ and $\mathcal{C} \not\subseteq \mathcal{S}$, and contain opposing values for at least one shared attribute. We select the five eligible recordings with the highest Jaccard similarity $|\mathcal{S} \cap \mathcal{C}|/|\mathcal{S} \cup \mathcal{C}|$ to obtain closely related but contrasting negatives.

Table 12 additionally ranks only the original caption or paired recording against the same five negatives by removing the other four positive candidates. Appendix C.4 defines Mean Rank and Top-1 accuracy (Acc@1) for the six remaining candidates.

Quality Control. Figures 4 and 5 present the speaking-style caption generation and Coco-Nut consensus caption templates, respectively. The speech-style prompt requests accurate variants that preserve the intended meaning and inaccurate captions that change one stated attribute while preserving the surrounding description. The Coco-Nut prompt consolidates Japanese listener descriptions into an English consensus caption. An example is shown in Figure 3. We primarily use Gemini 3.1 Flash-Lite to generate all structured caption data in this paper. We manually reviewed samples from every constructed dataset and samples of generated captions used in training and evaluation. Automatic validation checks the required output structure and ensures distinct caption candidates within each sample. When Gemini 3.1 Flash-Lite fails validation checks multiple times, we generate those captions with Gemini 3.5 Flash.

Table 7: Evaluation datasets and task definitions. N is the number of evaluation queries after task-specific selection, with audio-to-text / text-to-audio counts shown for retrieval.

Task	Dataset and evaluation partition	Annotation or candidate construction	N
Speech: Retrieval			
Audio ↔ Text	ParaSpeechCaps, official combined evaluation	One representative per tag set; one caption per clip	550 / 550
Sound: Retrieval			
Audio ↔ Text	AudioCaps, test	All available source captions	932 / 4,546
Audio ↔ Text	Clotho, evaluation	Five human captions per clip	1,045 / 5,225
Music: Retrieval			
Audio ↔ Text	MusicCaps, eval	Source music descriptions	2,730 / 2,730
Speech Classification: Speaking Style			
Gender	Common Voice 17.0, English test; age-gender balanced	2 classes; contributor gender metadata	168
Age	Common Voice 17.0, English test; same balanced subset	4 age groups; contributor age metadata	168
Accent	ESLTTS, derived evaluation partition from ada + test	17 accent classes; train/eval speaker-disjoint	2,773
Pitch	VCTK, filtered and balanced utterances	6 classes: gender × low/medium/high pitch; PENN F0 estimates	246
Speaking Rate	VCTK, filtered and balanced utterances	3 classes; estimated syllables per second of detected speech	300
Fluency	SpeechOcean762, capped test subset	4 bins of human fluency scores: 0-3, 4-5, 6-7, 8-10; at most 60 clips per class	199
Emotion	CREMA-D, all prepared clips	6 acted-emotion labels encoded in source filenames	7,442
Emotion	MSP-Podcast 2.0, balanced Test2	8 emotions; human-vote selection ($\geq$5 valid ratings; top-vote fraction $\geq$0.5)	1,000
Speech Classification: Recording Conditions			
Distance	VCTK convolved with simulated room impulse responses	3 controlled source-microphone distance classes: close-talk, near-field, far-field	300
Reverb Level	VCTK with controlled reverberation	3 classes: dry, mildly reverberant, strongly reverberant; loudness normalized	300
Speech Classification: Conversation			
Overlap Detection	AMI, test; 8-second crops	no-overlap / overlap $\geq$0.2 seconds	200
Turn Counting	AMI, test; 12-second crops	4 classes (1 to 4 turns); backchannels excluded	400
Dialog Act	DailyTalk, val; shared speakers	4 source labels: commissive, directive, inform, question	1,197
Backchanneling	CANDOR	2 classes; manually annotated backchannels	270
Speech Classification: Speech Content			
Keyword Spotting	Speech Commands v0.02, test	30 spoken-word classes	10,187
Intent Classification	Fluent Speech Commands, test	31 action-object-location intents	3,793
Sound: Classification			
Environmental Sound	ESC-50, all five folds	50 source sound-event labels	2,000
Paralinguistic Events	VocalSound, test	cough, laughter, sigh, throat clearing, sneeze, sniff	3,591
Music: Classification			
Musical Instrument	NSynth, test	10 instrument family labels	4,096
Music Genre	GTZAN	10 source genre labels; corrupt audio excluded	999
Speech: Ranking			
Audio → Text	ParaSpeechCaps, balanced holdout	Speech-style caption plus generated accurate and inaccurate variants	994
Text → Audio	ParaSpeechCaps, balanced holdout	Original caption; accurate and contrasting recordings selected by style tags	789
Audio → Text	Coco-Nut, Japanese test audio	English consensus caption synthesized from Japanese listener descriptions	578
Sound: ranking			
Audio → Text	AudioCaps, test; available validated-caption subset	Sound caption plus generated structured variants	913
Music: ranking			
Audio → Text	MusicCaps, eval	Music caption plus generated structured variants	2,730

```
You generate contrastive speech-style captions for audio-text representation learning.

You are given:
- `ground_truth_caption`: the primary natural-language caption describing the speech/audio. Treat this as the
base text to rewrite or minimally edit.
- `atomic_attributes`: explicit metadata facts used for grounding, validation, and possible single-attribute
flips. They are not a checklist of details that every generated caption must contain.
- `attribute_context`: a structured dictionary of the same or related attributes, provided only as extra audit
context.

Your task is to generate hard positive and hard negative captions.

Definitions:
- A hard positive caption must be semantically equivalent to, or entailed by, the ground-truth caption.
- A hard positive may reorder information, paraphrase words, change syntax, use label-style formatting, or omit
some details for a partial positive.
- A hard positive must not introduce unsupported new attributes or expand the caption by listing metadata-only
attributes that were absent from the ground-truth caption.
- A hard negative caption must be a minimal edit of the ground-truth caption that is clearly contradictory.
- Each hard negative must change exactly one attribute expressed in the ground-truth caption while keeping all
other caption content unchanged or closely paraphrased.
- Omission alone is not a valid negative. A negative must explicitly contradict one ground-truth attribute.
- Preserve modifiers such as "slightly", "occasionally", "smoothly", "measured", and "clean" unless that
modifier is the changed attribute.
- Do not add extra clauses from `atomic_attributes` just because they are present in metadata.
- If an atomic attribute conflicts with the ground-truth caption, do not use that conflict to rewrite the
caption wholesale. Keep the generated caption anchored to the ground-truth caption and change at most one stated
attribute for a negative.
- For negative metadata, choose `original_value` values that are supported by the ground-truth caption and use
the exact wording from `atomic_attributes` whenever an equivalent attribute appears there.
- For gender flips, prefer metadata wording such as `original_value: "female"` and `new_value: "male"` when
`atomic_attributes` uses `female`/`male`, even if the caption prose says "woman" or "man".
- Prefer categorical or near-neighbor substitutions for hard negatives, such as clean → noisy, high-pitched →
low-pitched, silky → rough, measured speed → rapid speed, female → male.
- Avoid weak modifier-only negatives such as clean → very clean, noisy → slightly noisy, high-pitched → very
high-pitched, or silky → very silky. These often remain semantically compatible with the original caption.
- Do not use `modifier_flip` hard negatives unless the modifier change creates a clear categorical
contradiction. Prefer `single_attribute_flip`, `near_neighbor_substitution`, or `frequency_flip` when possible.

Generate exactly:
- 10 hard positives
- 10 hard negatives

Each of the five transformation types below should be used exactly twice to generate 10 hard positives:
- `reordering`: Change the order of information while preserving all facts.
- `paraphrase`: Replace words with semantically equivalent or near-equivalent phrases.
- `syntax`: Convert one long caption into multiple shorter clauses, or vice versa.
- `label_style`: Generate captions that sound like dataset annotations rather than natural prose.
- `partial_positive`: A positive caption may omit some details if the remaining details are still true.

The 10 hard negatives should cover diverse attribute types when possible, such as:
- speaker gender
- accent
- speaking rate
- pitch
- voice quality
- fluency
- background noise
- recording condition
- paralinguistic event
- categorical presence/absence changes
- near-neighbor substitutions
- event frequency changes only when the caption contains an event frequency

Do not include explanations outside the JSON.

Example input:

<input_json>
{EXAMPLE_INPUT_JSON}
</input_json>

Example output:

{EXAMPLE_OUTPUT_JSON}

Now generate captions for this input:

<input_json>

{PER_RECORD_INPUT_JSON}

</input_json>
```

Figure 4: Prompt template for generating accurate and inaccurate speech-style captions. The prompt requests accurate rewrites that preserve or omit supported details and inaccurate edits that contradict one stated attribute while preserving the surrounding content. Accurate and inaccurate captions are constructed to preserve or contradict the reference description, subject to annotation noise and sampled manual review. Syntax variants are discarded during training and evaluation.

```
You generate contrastive English speech-style captions from Japanese listener annotations.

You are given multiple Japanese descriptions of the same utterance:
- `ground_truth_caption`: a compact preprocessing note, not a caption to translate or copy.
- `atomic_attributes`: the Japanese listener descriptions. For the test split there are usually five descriptions from different
listeners.
- `attribute_context`: lightweight metadata such as language and split for audit context.

Your task is to first synthesize one faithful English perceptual reference caption, then generate hard positive and hard negative
captions.

Important: this is consensus-aware synthesis, not sentence-by-sentence translation. The Japanese descriptions may disagree because they
were written by different listeners. Do not assume every detail is correct. Extract the perceptual attributes from each description,
identify shared patterns and disagreements across annotators, and synthesize them into one coherent English caption.

Consensus and ambiguity rules:
- If most annotators agree on an attribute, state it directly.
- If only some annotators agree, use softened wording such as "appears to", "is perceived as", "sounds somewhat", or "mostly".
- If descriptions strongly contradict each other, mention the ambiguity naturally, for example "the speaker's gender is somewhat
ambiguous, though mostly perceived as male".
- Include a one-off detail only when it is salient and not contradicted by other descriptions.
- Age and gender are perceptual voice descriptions, not facts. Do not write exact demographic claims such as "a 25-year-old man".
Prefer "young adult male-sounding speaker", "speaker perceived as a young man", or "gender-ambiguous speaker with a youthful voice".

Caption scope:
- The top-level `caption` is the synthesized English reference caption for the utterance.
- Prioritize perceived age/gender, pitch, speaking rate, emotion or arousal, voice quality, speaking style, and special effects.
- Voice-quality examples include hoarse, raspy, nasal, breathy, synthetic, processed, distorted, robotic, artificially processed, or
voice-changer-like.
- Speaking-style examples include narration, live-commentary-like delivery, conversation, announcement, storytelling, excited delivery,
calm delivery, or fast-paced delivery.
- Do not hallucinate missing audio details. If descriptions say "live-commentary-like", do not invent a stadium, sports event, crowd,
background noise, or content topic.
- Do not use, quote, paraphrase, or depend on transcript content.
- Keep the top-level caption natural and concise, roughly 12 to 35 words. A good shape is: "A [perceived speaker type] with
[pitch/voice quality], speaking [rate] in a [style/emotion] manner."
- Write all generated captions in English.

Helpful Japanese-to-English phrasing:
- 男性 / 女性 disagreement: "perceived mostly as male", "gender-ambiguous", or "male-sounding".
- 若い / 20代から30代: "young adult" or "young or young-adult".
- 高い声: "high-pitched voice".
- だみ声: "raspy" or "hoarse".
- 機械で加工したような: "artificially processed".
- 変声機を使ったような: "voice-changer-like".
- 合成音声で加工したような: "synthetic or processed-sounding".
- 早口: "speaking quickly" or "fast-paced delivery".
- 興奮したように: "excited-sounding".
- 実況するような口調: "live-commentary-like delivery".

Definitions:
- A hard positive caption must be semantically equivalent to, or entailed by, the synthesized top-level `caption` and the consensus
interpretation of the Japanese descriptions.
- A hard positive may reorder information, paraphrase words, change syntax, use label-style formatting, or omit some non-dominant
details for a partial positive.
- A hard positive must preserve uncertainty and softened wording when the reference caption includes ambiguity.
- A hard negative caption must be a minimal, plausible edit that contradicts exactly one perceptual attribute from the synthesized
top-level caption.
- Negatives should change attributes such as perceived gender, perceived age group, pitch, speaking rate, emotion/arousal, voice
quality, speaking style, or processing/special effect.
- Use English `original_value` values supported by the synthesized top-level caption, and English `new_value` values that are plausible
but contradictory.
- Do not make a negative by adding unsupported setting, topic, background, transcript, room, microphone, recording-quality, or
source-language details.

Generate exactly:
- 1 top-level `caption`
- 10 hard positives
- 10 hard negatives

Each of the five transformation types below should be used exactly twice to generate 10 hard positives:
- `reordering`: Change the order of information while preserving all facts.
- `paraphrase`: Replace words with semantically equivalent or near-equivalent phrases.
- `syntax`: Split or combine clauses while preserving meaning.
- `label_style`: Generate captions that sound like dataset annotations rather than natural prose.
- `partial_positive`: Omit some non-dominant details while keeping the remaining caption true.

Return only valid JSON with the fields `caption`, `positives`, and `negatives`. Do not include explanations outside the JSON.

Example input:

<input_json>
{EXAMPLE_INPUT_JSON}
</input_json>

Example output:

{EXAMPLE_OUTPUT_JSON}

Now generate captions for this input:

<input_json>

{PER_RECORD_INPUT_JSON}

</input_json>
```

Figure 5: Prompt template for generating accurate and inaccurate speech-style captions from Japanese listener descriptions in Coco-Nut (Watanabe et al., 2023). Syntax variants are discarded during evaluation.

```
Create three complementary text-to-speech voice instructions from an attached reference recording.

You are given:
- one or two original captions describing the same recording;
- optional compact facts about the recording;
- the reference recording as attached audio.

Grounding rules:
- Use the reference audio as the primary evidence for audible voice and speaking-style attributes. Treat the original captions and source facts
as factual guardrails and useful priors, not as text to simply paraphrase.
- Derive all three instructions from the combined evidence. When the original caption is generic or incomplete, add only salient traits that are
clearly audible.
- Keep APS, DSD, and RP mutually consistent with one another and with the reference audio plus explicit text facts; do not introduce
contradictions across the three captions.
- Include a detail only when it is audible in the recording or directly supported by the provided text. Do not guess a precise age, accent,
gender presentation, emotion, or other trait from weak evidence; use a broad or explicitly uncertain description when necessary.
- When clear audio evidence conflicts with an original caption about an audible trait, prefer the audio. When the audio is ambiguous or seems to
conflict with an explicit source fact, keep the source fact and soften the audio inference with wording such as "appears", "likely", or
"somewhat".
- Do not invent identities, backstory, transcript content, environment, relationships, medical conditions, or emotions that are not grounded in
the voice style.

Create these three instruction styles:
- APS (Acoustic-Parameter Specification) is a structured attribute sheet for fine-grained control of the voice. Fill all twelve fields below.
Each value should describe the named acoustic, vocal, emotional, or speaker-presentation dimension directly rather than instructing the reader
how to perform it.
- DSD (Descriptive-Style Directive) is one coherent performance instruction that tells a performer or speech synthesizer how the voice should
sound. Select around 5 distinct, salient attributes from the full description, give striking traits priority, and weave them into natural prose
instead of listing fields. This is a voice-style instruction, not a character or scenario.
- RP (Role-Play) is a director's note built around a specific character persona inferred from the same voice style. Describe the character and
manner of speaking in a vivid, colloquial, immediately actionable phrase. The persona may creatively summarize the audible style, but it must not
assert a real identity or unsupported backstory.

APS field definitions:
- gender: perceived voice gender or gender presentation, phrased as voice style rather than identity when uncertain.
- pitch: perceived high/low frequency, range, stability, and notable pitch movement relative to the speaker presentation.
- speed: speaking rate, rhythm, pauses, and any clear acceleration or slowdown over the clip.
- volume: loudness and intensity, from whispering or soft speech through conversational volume to raised or shouted delivery.
- age: broad apparent life stage inferred from voice, such as child, teenager, young adult, adult, middle-aged adult, or elderly.
- clarity: pronunciation and articulation quality, including crisp, precise, mumbled, slurred, blurred, or run-together speech.
- fluency: smoothness and continuity, including hesitations, repetitions, fillers, breaks, or especially effortless flow.
- accent: pronunciation variety or dialect; name a specific region only when clearly supported, otherwise use a broad variety such as American
English, British English, or Mandarin Chinese.
- texture: voice timbre, such as warm, bright, husky, nasal, breathy, gravelly, deep, mellow, delicate, or resonant.
- emotion: audible feeling or affect, including shifts such as calm becoming frustrated or sadness giving way to laughter.
- tone: attitude conveyed by inflection, such as formal, sarcastic, enthusiastic, detached, reassuring, urgent, or matter-of-fact.
- personality: prominent, consistent speaker presence suggested by the voice, such as confident, reserved, assertive, anxious, playful, or
composed.

APS writing rules:
- Keep every APS value as a concise phrase: more informative than a bare label, shorter than a full explanation.
- For pitch, speed, and volume, describe separate evidence; do not label all of them "moderate" unless the audio truly gives no stronger cue.
- Note clear changes over time in the relevant field instead of forcing a static description.

DSD writing rules:
- Write one or two fluent declarative or imperative sentences in a natural monologue style, never dialogue or questions.
- Keep the result idiomatic and instantly actionable, not a comma-separated field dump.
- Do not start DSD with "Speak", "Deliver", "Use", "Adopt", "Project", "Maintain", "Channel", "Utilize", "Imagine", or "Envision".

RP writing rules:
- Include a concrete character role or persona and integrate the manner of speaking into the same director's note.
- Use declarative sentences only. Do not write dialogue, quotations, or questions.
- Do not start RP with framing words such as "Like", "Imagine", "Pretend", "As", "You are", or "Acting"; do not write actual dialogue lines.

General writing rules:
- Vary sentence structures naturally across samples; avoid fixed templates.
- Do not mention dataset names, model names, prompts, file paths, or evaluation setup.
- Return only valid JSON. Do not include explanations outside the JSON.

Output exactly this JSON shape:
{
  "aps": {
    "gender": "",
    "pitch": "",
    "speed": "",
    "volume": "",
    "age": "",
    "clarity": "",
    "fluency": "",
    "accent": "",
    "texture": "",
    "emotion": "",
    "tone": "",
    "personality": ""
  },
  "dsd": "",
  "rp": ""
}

Example input:

<input_json>
{EXAMPLE_INPUT_JSON}
</input_json>

Example output:

{EXAMPLE_OUTPUT_JSON}

Now generate APS, DSD, and RP instruction captions for this input.

<input_json>
{PER_RECORD_INPUT_JSON}
</input_json>

The reference audio is attached as inline audio after this text prompt.
```

Figure 6: Prompt template for generating instruction captions: APS, DSD, and RP. The prompt follows a similar principle to InstructTTSEval (Huang et al., 2025).

B FINE-TUNING SETUP

B.1 TRAINING DATA

The training corpus combines 17 speech, sound, and music sources with the counts, partitions, and sampling weights in Table 8. We exclude evaluation recordings from the fine-tuning corpus. DailyTalk contains two speakers and uses recording-disjoint training and validation partitions with shared speakers (Lee et al., 2023). The corpus connects speech style, recording conditions, conversation, and content with complementary sound and music supervision.

For ParaSpeechCaps, we additionally generate three instruction-caption families to connect speech representation learning with TTS adherence evaluation in the forms used by InstructTTSEval (Huang et al., 2025). Gemini 3.1 Flash-Lite uses each ParaSpeechCaps recording as the primary evidence, together with the original caption and available attributes, to produce mutually consistent training captions in the APS, DSD, and RP families. These generated training captions are distinct from the InstructTTSEval evaluation instructions. A second stage constructs accurate variants and targeted contradictory captions for each family, using structured attribute edits for APS and instruction-preserving transformations for DSD and RP. The resulting captions teach the same audio-text model to compare recordings with descriptive and instruction-style language. Appendix A.3 describes the shared manual review and automatic validation. Figures 4 and 6 present the speech-style caption-generation template and the instruction captions for ParaSpeechCaps, respectively.

B.2 MULTI-POSITIVE/NEGATIVE SIGLIP

For completeness, the positive and negative pair sets used by the standard contrastive term are

$$\mathcal{P}^+ = \bigcup_{i=1}^{B} \left\{ (\boldsymbol{a}_i, \boldsymbol{t}^+) : \boldsymbol{t}^+ \in \mathcal{T}_i^+ \right\}, \tag{8}$$

$$\widehat{\mathcal{T}}_i^- = \mathcal{T}_i^- \cup \bigcup_{j \neq i} \left(\mathcal{T}_j^+ \cup \mathcal{T}_j^- \right), \qquad \mathcal{P}^- = \bigcup_{i=1}^{B} \left\{ (\boldsymbol{a}_i, \boldsymbol{t}^-) : \boldsymbol{t}^- \in \widehat{\mathcal{T}}_i^- \right\}. \tag{9}$$

Here $\mathcal{T}_i^+$ includes all accurate captions for recording i, including captions shared by multiple recordings. All known positive and ignored relations are removed from $\mathcal{P}^-$ before the loss is evaluated, so $\mathcal{P}^+ \cap \mathcal{P}^- = \varnothing$. Known positive pairs and eligible negative pairs, including inaccurate captions and in-batch contrasts, receive equal weight. The shared calibrated score is

$$s_{\boldsymbol{a},\boldsymbol{t}} = \alpha\, \boldsymbol{a}^\top \boldsymbol{t} + b, \tag{10}$$

where α and b are the shared trainable scale and scalar bias. The resulting pair-mean loss is

$$\mathcal{L}_{\text{SigLIP}} = -\frac{\sum_{(\boldsymbol{a},\boldsymbol{t}^+) \in \mathcal{P}^+} \log \sigma(s_{\boldsymbol{a},\boldsymbol{t}^+}) + \sum_{(\boldsymbol{a},\boldsymbol{t}^-) \in \mathcal{P}^-} \log \sigma(-s_{\boldsymbol{a},\boldsymbol{t}^-})}{|\mathcal{P}^+| + |\mathcal{P}^-|}, \tag{11}$$

where σ is the sigmoid function. Both SigLIP and SupCon losses use the shared score calibration, and the scalar bias cancels in the SupCon softmax. VoiceCLAP S and PE_{AV} retain raw embeddings, with PE_{AV} using the backbone's native audio-text calibration. M2D-CLAP normalizes embeddings (cosine similarity), initializing $\alpha = 1$ and $b = -10$. All SigLIP ablations use the same pair-mean normalization.

B.3 HYPERPARAMETERS

Optimization. The default recipe fully fine-tunes the audio and text encoders in pretrained audio-text models. We freeze the DAC-VAE audio feature extractor and remove the video and audio-video fusion encoders in PE_{AV} models. We fine-tuned all models on four RTX A6000 GPUs. By default, each GPU holds 20 audio recordings and their corresponding structured captions, resulting in a global batch size $B = 80$. Notably, because PE_{AV}L has a 1.1B-parameter audio encoder, we reduce the global batch size to $B = 48$ and introduce gradient accumulation for every two steps. AdamW uses a learning rate of 5×10^{-5} and weight decay of 0.01. Training follows a linear schedule with 200 warmup steps and 8,000 total updates. SigLIP and bidirectional SupCon have equal weights, giving $\lambda = 1$. The default seed is 0, with other seeds examined in Appendix C.6.

Caption sampling and mining. Each anchor samples up to $N_T^+ = 4$ accurate captions and $N_T^- = 16$ inaccurate captions. Accurate caption types include source ground-truth captions, reordered and paraphrased variants, label-style descriptions, and partial positives. Caption families include original descriptions and APS, DSD, and RP instructions, according to source availability. The miner selects up to $N_A^- = 8$ recordings by highest mean negative-caption-to-audio similarity, as in Eq. 1. Known positive recordings are excluded from mining.

B.4 Baseline Models

Tables 9 to 13 compare complementary audio-text models and audio LLMs. General audio-text comparators include MS-CLAP[23'] (Elizalde et al., 2024),[3] LAION-CLAP (Wu et al., 2023),[4] Human-CLAP (Takano et al., 2025),[5] MGA-CLAP (Li et al., 2024),[6] and GLAP (Dinkel et al., 2026).[7] Speech-focused comparators include ParaCLAP (Jing et al., 2024),[8] the Combined variant of ParaSpeechCLAP (Diwan et al., 2026),[9] and CLSP (Yang et al., 2026).[10] VoiceCLAP S and VoiceCLAP L use LAION's released models (LAION, 2026).[11][12]

The larger representation models comprise AuroLA 3B and AuroLA 7B (Xu et al., 2026),[13][14] Omni-Embed-Audio (OEA) Nemotron-3B,[15] OEA Qwen-3B,[16] and OEA Qwen-7B (Yoo et al., 2026),[17] and LCO-Embedding 3B and LCO-Embedding 7B (Xiao et al., 2025).[18][19] Our five SLS adaptations use VoiceCLAP S, M2D-CLAP[25'] (Niizumi et al., 2025),[20] and PE_{AV}S,[21] PE_{AV}B,[22] and PE_{AV}L (Vyas et al., 2026).[23] Audio-LLM comparisons use Audio Flamingo Next (Ghosh et al., 2026),[24] Kimi-Audio (KimiTeam et al., 2025),[25] Qwen3-Omni (Xu et al., 2025),[26] and Gemini 3.7 Flash.[27]

[3] https://github.com/microsoft/CLAP
[4] https://hf.co/laion/clap-htsat-fused
[5] https://github.com/sarulab-speech/Human-CLAP
[6] https://github.com/Ming-er/MGA-CLAP
[7] https://hf.co/mispeech/GLAP
[8] https://github.com/KeiKinn/ParaCLAP
[9] https://hf.co/ajd12342/paraspeechclap-combined
[10] https://hf.co/yfyeung/CLSP
[11] https://hf.co/laion/voiceclap-small-v2
[12] https://hf.co/laion/voiceclap-large-v2
[13] https://hf.co/Jazzcharles/AuroLA-3B
[14] https://hf.co/Jazzcharles/AuroLA-7B
[15] https://hf.co/JudeJiwoo/OEA-Nemo3B-AC
[16] https://hf.co/JudeJiwoo/OEA-Qwen3B-AC
[17] https://hf.co/JudeJiwoo/OEA-Qwen6B-AC
[18] https://hf.co/LCO-Embedding/LCO-Embedding-Omni-3B
[19] https://hf.co/LCO-Embedding/LCO-Embedding-Omni-7B
[20] https://github.com/nttcslab/m2d
[21] https://hf.co/facebook/pe-av-small
[22] https://hf.co/facebook/pe-av-base
[23] https://hf.co/facebook/pe-av-large
[24] https://hf.co/nvidia/audio-flamingo-next-hf
[25] https://hf.co/moonshotai/Kimi-Audio-7B-Instruct
[26] https://hf.co/Qwen/Qwen3-Omni-30B-A3B-Instruct
[27] https://deepmind.google/models/model-cards/gemini-3-7-flash/

Table 8: Training corpus and supervision sources. Counts and durations describe the prepared anchor recordings before sampling and augmentation. Each anchor is associated with accurate and inaccurate caption pools. Relative source weights determine sampling.

Data source	Partition and supervision	Rows	Duration (hours)	Weight
Speech				
ParaSpeechCaps	train_base; source descriptions, generated caption variants, and APS/DSD/RP instruction captions	115,094	266.32	15
Common Voice 17.0	English train; retained gender, age, and accent annotations expressed as captions	582,412	912.03	1
ESLTTS	dev utterances from derived training-speaker partition; fixed accent captions	21,380	21.85	1
SpeechOcean762	train; human pronunciation/fluency scores, attribute captions, and generated variants	2,500	2.88	1
TextrolSpeech (emotional subset)	emotion_dataset recordings; generated captions	42,847	42.57	1
MSP-Podcast 2.0	filtered Train; emotion labels, speech attributes, and generated caption variants	102,629	153.96	1
AMI	train meetings; 12-second crops; joint speaker-count, turn-count, and overlap captions from timed annotations	551	1.84	1.5
CANDOR (backchanneling)	speaker-disjoint training conversations; filtered backchannels and matched listener-absence controls; fixed captions	47,208	85.87	1
CANDOR (turn counting)	same training partition; 12-second crops with 1 to 4 main turns; fixed captions	3,404	11.35	0.5
DailyTalk	train; dialog-act and speech-attribute captions with generated variants	22,479	20.51	1
Fluent Speech Commands	train; intent-label captions and generated variants	23,132	14.72	1
Speech Commands v0.02	train; fixed captions	84,843	23.11	1
Sound				
VocalSound	train; vocal-event labels expressed as captions	15,531	18.05	1
AudioCaps	train; evaluation-video IDs excluded across AudioCaps and MusicCaps; source captions and generated variants	44,750	124.31	1
Music				
MusicCaps	non-AudioSet-eval partition; same video-ID exclusions; source descriptions and generated variants	2,560	7.11	1
Music4All	full inventory filtered to exactly one mapped genre; fixed genre captions	65,513	545.94	1
NSynth	official train; fixed captions over 10 instrument families	289,205	321.34	1
Total (17 sources)	**Data source totals**	**1,466,038**	**2,573.74**	**31**

Table 9: Full retrieval results. $A \rightarrow T$ and $T \rightarrow A$ denote audio-to-text and text-to-audio retrieval, respectively. Recall@1, Recall@5, and Recall@10 are expressed as percentages. Each row highlighted in blue ("+ SLS") adapts the preceding pretrained backbone. **Bold** and underline denote the highest and second/third best scores in each column, respectively.

	Model Parameters		ParaSpeechCaps						AudioCaps						Clotho						MusicCaps					
			$A \rightarrow T$			$T \rightarrow A$			$A \rightarrow T$			$T \rightarrow A$			$A \rightarrow T$			$T \rightarrow A$			$A \rightarrow T$			$T \rightarrow A$		
Model	Audio	Text	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Random Guess	–	–	0.2	0.9	1.8	0.2	0.9	1.8	0.1	0.5	1.1	0.1	0.5	1.1	0.1	0.5	1.0	0.1	0.5	1.0	0.0	0.2	0.4	0.0	0.2	0.4
Audio-text Models																										
MS-CLAP[23*]	0.03B	0.1B	1.3	4.5	9.1	1.3	6.2	10.5	36.2	67.4	80.9	30.2	62.9	77.4	22.1	48.6	60.2	17.7	42.2	55.0	8.9	24.8	34.2	9.0	24.4	34.0
LAION-CLAP	0.03B	0.1B	1.1	4.0	7.6	0.4	2.7	5.6	39.4	73.0	85.7	28.8	62.6	77.4	18.9	46.5	60.9	17.1	41.3	54.9	3.8	13.0	19.2	4.0	12.6	20.3
Human-CLAP	0.03B	0.1B	0.5	3.8	5.5	0.4	2.4	5.3	33.6	63.1	78.0	26.4	60.6	75.9	18.9	43.3	56.3	15.0	38.5	51.7	4.2	13.4	19.3	5.1	14.9	21.8
MGA-CLAP	0.03B	0.1B	0.4	3.8	6.5	0.2	1.8	5.5	56.0	85.2	92.0	41.6	75.5	86.7	27.2	51.7	64.4	21.8	47.1	60.8	7.9	18.1	25.2	7.1	18.4	25.4
GLAP	0.09B	0.8B	0.7	2.9	6.5	0.7	3.5	7.8	52.9	82.6	90.9	37.8	70.9	83.2	20.7	47.8	61.4	19.3	44.4	58.1	13.7	31.4	42.9	12.7	31.0	41.4
ParaCLAP	0.2B	0.1B	0.9	2.5	5.3	0.2	1.8	5.3	0.1	0.5	0.6	0.1	0.7	1.2	0.1	1.1	1.6	0.2	0.6	1.3	0.1	0.3	0.4	0.0	0.2	0.6
ParaSpeechCLAP	0.3B	0.3B	38.0	**81.1**	90.9	38.9	**78.4**	90.0	0.1	0.6	1.7	0.0	0.8	2.4	0.1	0.5	0.8	0.1	0.8	1.1	0.0	0.3	0.5	0.1	0.5	0.8
CLSP	0.6B	0.1B	26.9	68.9	83.5	28.2	67.5	82.9	0.8	3.1	4.4	0.8	2.6	4.8	0.5	1.1	2.0	0.4	1.1	1.5	0.4	1.2	2.4	0.4	1.4	2.1
VoiceCLAP L	0.6B	7B	4.2	15.1	22.9	4.0	12.7	20.7	31.0	59.5	73.4	24.4	55.4	69.5	20.2	39.4	51.0	16.8	38.1	50.1	11.1	27.8	39.4	13.6	32.8	43.9
AuroLA 3B	0.6B	3B	2.9	12.2	20.2	2.7	10.5	18.2	61.6	86.3	92.9	46.3	78.2	88.3	35.5	60.5	73.2	**27.8**	53.8	**66.8**	22.5	44.7	57.5	23.3	48.4	61.0
AuroLA 7B	0.6B	7B	3.5	12.9	21.3	2.5	11.3	19.3	**64.3**	87.0	93.9	**47.9**	**79.1**	88.8	**36.8**	**62.9**	**73.5**	**27.8**	**54.1**	65.6	23.3	48.4	60.6	24.6	50.7	62.8
OEA Nemotron-3B	0.6B	3B	1.3	8.2	14.0	1.6	6.4	12.2	49.1	75.3	84.8	39.6	73.7	84.3	22.3	46.7	58.5	19.2	42.2	55.0	7.9	21.5	30.2	8.8	23.8	33.5
OEA Qwen-3B	0.6B	3B	2.7	8.7	16.5	2.2	7.3	11.5	45.9	76.8	84.9	38.5	71.5	84.1	23.3	44.2	57.0	18.7	42.3	55.4	7.6	22.1	31.0	10.4	24.4	33.1
OEA Qwen-7B	0.6B	7B	2.0	9.3	14.2	2.2	6.9	12.9	49.5	77.3	86.2	39.6	72.9	85.1	24.7	49.1	62.3	20.7	45.1	58.5	8.8	24.8	34.7	10.9	26.6	36.8
LCO 3B	0.6B	3B	1.5	5.3	8.4	1.3	4.5	8.0	29.1	57.5	71.1	23.2	50.8	64.9	22.6	43.4	54.6	18.9	41.9	53.8	9.4	24.5	34.7	10.5	24.9	36.0
LCO 7B	0.6B	7B	1.5	6.0	10.4	0.9	5.5	8.0	34.7	63.7	76.4	27.1	57.0	70.6	24.5	46.4	57.4	20.1	43.5	55.1	11.8	30.1	39.9	13.7	31.8	42.7
VoiceCLAP S	0.09B	0.02B	1.5	6.4	10.9	1.6	4.9	8.9	1.0	3.5	6.5	1.2	3.7	6.1	0.4	1.4	2.6	0.4	1.3	2.3	0.3	1.0	1.6	0.3	1.2	1.9
+ **SLS**	0.09B	0.02B	20.0	55.5	71.1	22.5	58.4	74.2	4.1	15.2	23.6	4.5	17.1	26.9	0.6	3.9	5.5	1.4	4.6	7.5	1.3	4.9	8.2	1.8	6.6	11.1
M2D-CLAP[25*]	0.09B	0.1B	0.9	3.6	7.3	0.7	3.3	6.2	59.8	84.9	92.4	42.5	77.3	**89.0**	21.1	46.5	60.6	17.6	43.0	55.4	11.6	28.4	38.5	11.7	29.0	40.0
+ **SLS**	0.09B	0.1B	22.0	61.3	75.5	22.5	62.9	78.5	37.1	68.9	81.1	29.4	67.4	82.2	12.4	31.2	42.3	11.9	31.8	45.0	10.2	25.6	34.9	12.7	31.8	43.5
PE_{AV}S	0.09B	0.3B	3.8	15.5	26.5	3.8	16.0	24.5	58.4	85.6	92.3	41.4	74.8	85.8	33.5	58.5	71.8	23.9	50.9	63.3	24.2	49.0	60.7	25.0	49.5	61.4
+ **SLS**	0.09B	0.3B	32.4	73.6	86.4	34.9	74.0	87.6	52.6	81.7	89.7	35.6	70.6	82.6	27.8	51.4	64.0	19.5	45.6	59.4	20.1	45.6	56.6	20.1	42.9	54.5
PE_{AV}B	0.2B	0.3B	4.9	16.9	22.9	3.6	15.8	24.7	58.2	85.6	93.9	42.1	75.3	86.5	33.7	60.2	72.0	24.8	50.9	63.4	26.0	52.5	64.1	26.7	53.8	65.1
+ **SLS**	0.2B	0.3B	37.5	77.5	90.9	40.4	77.8	**91.6**	54.7	84.1	91.8	37.3	72.7	84.6	30.1	54.8	65.8	22.0	48.2	61.5	23.6	48.1	60.1	21.9	45.2	56.8
PE_{AV}L	1B	0.3B	5.1	15.6	25.3	4.5	17.1	25.6	63.9	**87.8**	**94.2**	45.6	77.4	87.7	34.4	59.4	71.8	24.6	51.4	63.5	**28.6**	**54.3**	**66.2**	**30.0**	**56.0**	**66.2**
+ **SLS**	1B	0.3B	**39.5**	78.0	**91.1**	**40.9**	**78.4**	90.9	57.2	85.2	92.9	39.8	75.6	86.4	29.8	53.0	65.9	22.4	48.3	61.1	24.3	49.1	60.5	21.9	47.0	57.8

C Full Results

This section reports complete results of all comparators, pretrained backbones, and SLS-adapted audio-text models on retrieval (Appendix C.1), classification (Appendix C.2), ten-candidate ranking (Appendix C.3), ranking ground truth against negatives (Appendix C.4), and reviewed same-speaker speaking-style contrasts (Appendix C.5).

C.1 Retrieval

Table 9 reports full retrieval results. All five SLS adaptations improve ParaSpeechCaps retrieval in both directions. The adapted PE_{AV} variants exceed CLSP at every reported recall cutoff, and adapted PE_{AV}L leads Recall@1 in both directions. Even adapted VoiceCLAP S exceeds the larger VoiceCLAP L, AuroLA, OEA, and LCO variants on speech retrieval. ParaSpeechCLAP retains the highest audio-to-text Recall@5 and ties adapted PE_{AV}L on text-to-audio Recall@5. These comparisons show that SLS strengthens speech correspondence across three backbone families.

Sound and music retrieval reveal complementary domain strengths because of the introduction of additional sound and music data into the training recipe. VoiceCLAP S improves across all four retrieval datasets, while the adapted PE_{AV} models retain substantial sound and music retrieval capabilities alongside their speech gains. Changes for M2D-CLAP and PE_{AV} vary by dataset and direction, with AuroLA and several pretrained backbones remaining strong on sound and music. The full results distinguish SLS's consistent speech gains from backbone-dependent behavior in other domains.

Table 10: Full classification results. Values are task-level macro F1 scores on a 0 to 100 scale. Each row highlighted in blue ("+ SLS") adapts the preceding pretrained backbone. **Bold** and underline denote the highest and second/third best scores in each column, respectively.

	Model Parameters		Speaking Style								Recording Conditions		Conversation				Content		Sound		Music	
			CV17 Gender	CV17 Age	ESLTTS Accent	VCTK Pitch	VCTK Speaking Rate	speechocean762 Fluency	CREMA-D Emotion	MSP-Podcast Emotion	VCTK Distance	VCTK Reverberation	AMI Overlap	AMI Turn Count	DailyTalk Dialog Act	CANDOR Backchannel	Speech Commands Keyword	Fluent Speech Intent	ESC-50	VocalSound	NSynth Instrument	GTZAN Genre
Model	Audio	Text																				
Random Guess	–	–	50.0	25.0	5.9	16.7	33.3	25.0	16.7	12.5	33.3	33.3	50.0	25.0	25.0	50.0	3.3	3.2	2.0	16.7	10.0	10.0
Audio-text Models																						
MS-CLAP[23']	0.03B	0.1B	93.4	28.0	9.9	34.5	28.8	21.1	25.4	14.9	42.4	30.5	65.0	23.7	20.9	38.9	4.4	1.6	**95.6**	77.6	66.8	67.4
LAION-CLAP	0.03B	0.1B	89.9	18.3	4.1	36.8	35.6	25.3	19.0	8.8	46.0	79.2	53.3	24.9	17.1	41.3	4.8	2.5	92.9	79.9	31.2	45.2
Human-CLAP	0.03B	0.1B	91.6	27.0	3.6	31.0	29.9	25.5	17.4	12.1	35.3	66.3	52.0	19.8	15.7	50.0	4.3	2.5	91.5	80.2	32.9	51.1
MGA-CLAP	0.03B	0.1B	93.4	26.5	4.2	38.1	32.3	18.2	28.0	15.3	44.4	53.8	62.4	21.8	21.5	49.7	3.5	1.2	95.5	82.5	37.9	52.5
GLAP	0.09B	0.8B	80.6	32.4	4.5	25.3	40.6	18.6	12.2	9.6	31.1	38.0	51.0	15.1	20.5	60.3	95.4	39.1	91.5	79.2	27.8	66.6
ParaCLAP	0.2B	0.1B	43.1	19.9	4.0	14.3	20.8	25.0	21.7	21.7	25.7	20.3	33.3	12.5	5.0	33.2	2.7	1.2	0.3	14.3	7.8	3.1
ParaSpeechCLAP	0.3B	0.3B	90.5	26.5	7.5	41.2	47.8	29.6	19.1	20.0	38.5	27.5	50.9	21.2	11.2	48.3	2.2	2.0	1.0	10.8	6.6	9.3
CLSP	0.6B	0.1B	96.4	46.9	12.5	42.5	58.4	34.9	31.5	21.9	29.5	37.3	55.0	26.9	41.3	42.1	9.3	2.2	2.4	38.3	8.9	13.0
VoiceCLAP L	0.6B	7B	**97.6**	**51.1**	23.5	33.6	37.5	38.0	48.8	**34.5**	33.5	48.7	66.1	28.6	51.0	60.6	87.4	72.0	79.7	88.5	39.7	80.6
AuroLA 3B	0.6B	3B	97.0	28.7	12.5	39.6	37.7	34.2	35.1	29.4	37.4	44.4	72.8	29.8	48.5	59.8	97.0	95.2	91.4	92.8	41.9	79.0
AuroLA 7B	0.6B	7B	96.4	36.5	11.8	41.4	40.2	34.6	49.3	31.0	34.2	43.6	75.5	33.0	52.4	56.2	**97.1**	93.6	92.6	93.4	49.4	79.7
OEA Nemotron-3B	0.6B	3B	94.6	27.7	2.9	37.3	28.7	29.2	31.1	17.4	51.6	43.6	79.5	**39.8**	29.9	45.0	55.2	51.3	88.0	89.6	33.4	55.3
OEA Qwen-3B	0.6B	3B	95.8	29.7	3.5	32.1	31.9	27.8	39.2	21.1	34.0	35.4	71.0	28.5	34.1	75.0	44.3	36.2	85.4	89.0	31.5	59.1
OEA Qwen-7B	0.6B	7B	89.3	22.9	6.4	30.7	29.9	26.1	43.4	17.0	32.4	28.7	73.5	36.8	25.0	38.8	37.7	54.4	88.9	91.3	35.8	57.5
LCO 3B	0.6B	3B	58.2	24.3	7.4	21.6	27.4	27.7	33.8	24.4	33.2	32.8	50.5	22.3	49.0	46.7	96.5	97.1	88.1	89.5	51.1	77.5
LCO 7B	0.6B	7B	60.6	22.4	6.0	26.3	30.5	24.9	45.6	25.3	34.9	34.9	57.4	26.8	47.3	47.4	96.6	98.3	88.1	92.7	54.5	78.8
VoiceCLAP S	0.09B	0.02B	91.7	39.8	6.4	41.0	55.3	16.2	19.1	16.7	16.6	25.0	43.9	17.7	32.6	44.6	5.4	1.4	2.9	46.5	3.9	2.2
+ **SLS**	0.09B	0.02B	93.4	44.7	36.5	54.1	60.3	**58.6**	28.2	21.2	91.6	37.0	48.6	27.8	54.6	93.7	91.5	66.8	14.0	90.8	62.0	22.9
M2D-CLAP[25']	0.09B	0.1B	93.5	23.3	6.9	37.5	39.4	25.0	23.3	15.6	43.4	48.4	52.1	23.8	14.8	51.0	3.2	2.9	90.3	82.5	38.9	71.1
+ **SLS**	0.09B	0.1B	94.0	46.2	32.9	50.6	46.7	42.4	29.1	19.6	82.1	52.7	69.2	36.4	45.0	94.4	95.8	76.7	82.8	94.0	**77.5**	73.7
PE_{AV}S	0.09B	0.3B	95.8	44.2	18.2	**68.7**	**65.2**	18.2	41.7	22.7	50.7	52.2	56.3	27.3	28.4	65.8	69.4	23.6	93.5	81.8	72.0	76.6
+ **SLS**	0.09B	0.3B	95.8	45.0	42.4	60.6	63.9	52.5	37.7	23.7	**92.3**	80.0	78.3	35.6	51.7	94.8	94.3	89.5	88.2	93.9	75.4	73.1
PE_{AV}B	0.2B	0.3B	96.4	45.6	16.3	65.1	61.8	21.9	40.1	23.2	50.4	45.7	54.9	24.6	25.3	50.3	72.6	25.9	94.6	82.2	67.7	76.5
+ **SLS**	0.2B	0.3B	92.8	47.8	**44.1**	59.8	64.5	45.2	37.5	26.5	81.1	**84.7**	75.3	34.2	54.0	94.4	94.8	94.3	86.8	93.7	76.1	72.5
PE_{AV}L	1B	0.3B	95.2	48.4	15.4	68.3	63.9	29.0	41.5	23.6	57.4	60.4	57.8	31.5	25.9	51.8	71.7	28.9	95.0	84.4	69.6	75.1
+ **SLS**	1B	0.3B	94.0	48.7	36.0	58.3	62.9	44.0	35.6	26.8	87.9	83.0	79.8	34.4	62.0	**96.7**	96.2	96.2	89.6	**94.3**	74.0	78.5
Audio LLMs																						
AF-Next	0.6B	7B	57.5	27.3	2.1	25.1	16.7	27.5	**55.9**	27.6	24.2	27.4	65.0	19.9	49.9	48.9	94.7	78.0	84.7	89.9	72.8	80.0
Kimi-Audio	0.6B	7B	88.5	35.0	7.4	25.2	38.2	21.2	35.3	25.3	37.7	33.6	39.4	20.4	46.2	49.9	94.1	97.9	73.8	92.1	21.1	57.8
Qwen3-Omni	0.6B	30B	95.8	41.5	12.9	35.1	33.0	28.8	25.8	33.4	19.3	34.9	75.5	31.1	59.9	84.0	96.7	97.5	83.4	91.8	30.1	**89.1**
Gemini 3.7 Flash	–	–	89.2	44.5	36.4	42.5	44.9	29.2	25.4	27.1	47.0	59.7	**85.9**	34.7	**63.1**	82.8	96.9	**99.5**	61.0	86.2	16.9	75.9

C.2 CLASSIFICATION

Table 10 reports full classification results, including speech, sound, and music domain tasks. The results highlight task-level strengths across the four speech families. Adapted PE_{AV}B leads accent and reverberation classification, adapted PE_{AV}S leads distance classification, and adapted VoiceCLAP S leads fluency classification. All five SLS adaptations exceed the listed unadapted models and audio LLMs on backchannel detection. Moreover, the SLS-adapted VoiceCLAP S exceeds VoiceCLAP L on accent, pitch, rate, fluency, distance, and backchannels. At the family level, Table 1 shows that all five adaptations exceed CLSP and ParaSpeechCLAP in each of the four classification-family averages. Complementary sound and music tasks retain distinct backbone strengths. Adapted M2D-CLAP leads instrument classification, and adapted PE_{AV}L leads vocal-event classification. These results connect SLS's fine-grained supervision with useful discrimination beyond speech.

C.3 RANKING

Table 11 reports ten-candidate ranking with five positive candidates and five negatives, using PairAcc and strict separation (StrictSep). StrictSep is the proportion of queries for which every positive

Table 11: Full ranking results. Caption ranking ranks captions for a recording, and audio ranking ranks recordings for a caption. Each row highlighted in blue ("+ SLS") adapts the preceding pretrained backbone. **Bold** and underline denote the highest and second/third best scores in each column, respectively.

	Model Parameters		ParaSpeechCaps				Coco-Nut		AudioCaps		MusicCaps	
			Caption		Audio		Caption		Caption		Caption	
Model	Audio	Text	PairAcc	StrictSep	PairAcc	StrictSep	PairAcc	StrictSep	PairAcc	StrictSep	PairAcc	StrictSep
Random Guess	–	–	50.0	0.4	50.0	0.4	50.0	0.4	50.0	0.4	50.0	0.4
Audio-text Models												
MS-CLAP[23']	0.03B	0.1B	41.9	0.0	54.3	6.5	57.3	1.4	97.9	82.5	69.1	4.8
LAION-CLAP	0.03B	0.1B	47.6	0.5	51.9	2.2	59.6	2.6	97.6	78.5	56.3	2.2
Human-CLAP	0.03B	0.1B	45.1	0.0	49.5	2.8	60.0	1.0	99.1	92.7	52.7	2.1
MGA-CLAP	0.03B	0.1B	57.6	0.6	49.4	4.1	64.3	3.6	**99.9**	98.5	64.3	1.8
GLAP	0.09B	0.8B	71.3	2.0	53.1	1.9	59.8	1.4	99.5	95.0	67.0	4.4
ParaCLAP	0.2B	0.1B	40.2	0.0	47.5	3.5	53.0	0.7	48.4	2.1	53.8	1.6
ParaSpeechCLAP	0.3B	0.3B	63.4	2.2	**81.5**	44.7	56.9	0.7	63.7	7.9	43.6	0.3
CLSP	0.6B	0.1B	64.7	1.0	**81.5**	48.3	73.1	7.4	62.4	14.5	47.3	0.7
VoiceCLAP L	0.6B	7B	70.4	4.1	65.0	13.1	76.8	11.9	99.6	95.9	52.0	2.5
AuroLA 3B	0.6B	3B	72.2	2.8	66.3	13.9	74.8	7.8	99.8	97.4	73.9	10.3
AuroLA 7B	0.6B	7B	69.1	2.7	67.7	13.7	75.6	10.2	99.7	96.9	73.6	10.9
OEA Nemotron-3B	0.6B	3B	66.7	3.0	56.6	6.0	69.4	5.2	99.6	97.2	54.8	2.0
OEA Qwen-3B	0.6B	3B	64.8	0.7	56.6	4.1	70.6	6.7	99.7	97.2	51.9	1.9
OEA Qwen-7B	0.6B	7B	66.4	2.1	57.1	3.7	71.1	7.6	99.7	96.8	52.8	1.5
LCO 3B	0.6B	3B	54.5	0.7	55.9	2.4	67.2	7.6	99.7	96.6	62.5	6.0
LCO 7B	0.6B	7B	59.7	1.4	56.0	1.4	67.9	5.4	99.8	97.9	59.1	5.3
VoiceCLAP S	0.09B	0.02B	70.8	3.2	60.1	8.4	57.4	2.6	69.5	11.5	40.5	1.0
+ **SLS**	0.09B	0.02B	96.8	71.4	79.4	37.5	71.3	13.3	90.0	46.3	94.7	46.9
M2D-CLAP[25']	0.09B	0.1B	56.4	0.2	55.8	5.2	62.1	2.6	**99.9**	97.8	60.9	3.7
+ **SLS**	0.09B	0.1B	96.5	70.1	80.5	43.7	78.3	18.3	96.3	71.7	97.5	70.3
$PE_{AV}S$	0.09B	0.3B	65.1	1.9	64.8	11.2	71.2	7.6	99.8	96.4	59.7	3.2
+ **SLS**	0.09B	0.3B	97.3	72.3	**81.5**	46.8	75.9	13.8	99.7	96.5	97.3	69.9
$PE_{AV}B$	0.2B	0.3B	67.0	2.1	62.0	9.4	75.2	6.6	99.7	95.5	60.2	3.1
+ **SLS**	0.2B	0.3B	97.3	72.8	79.1	46.1	75.2	11.9	99.6	95.8	97.6	72.5
$PE_{AV}L$	1B	0.3B	61.5	1.1	65.2	10.0	68.8	6.6	99.6	94.9	57.8	2.3
+ **SLS**	1B	0.3B	97.7	76.6	77.7	44.9	69.8	10.9	99.1	90.9	97.8	73.9
Audio LLMs												
AF-Next	0.6B	7B	69.5	6.7	52.4	0.5	78.8	19.9	95.1	70.2	77.4	18.0
Kimi-Audio	0.6B	7B	79.6	16.5	60.3	4.4	85.6	31.1	94.8	67.5	82.9	23.0
Qwen3-Omni	0.6B	30B	85.1	27.5	61.7	7.0	93.7	61.9	97.9	88.6	90.2	44.0
Gemini 3.7 Flash	–	–	**99.8**	**98.6**	77.9	**57.5**	**99.8**	**96.9**	99.7	**99.2**	**99.9**	**99.3**

candidate outranks every negative candidate. The per-query indicator is

$$\text{StrictSep}(\boldsymbol{q}) = \mathbb{1}\left[\min_{\boldsymbol{p}\in\mathcal{P}(\boldsymbol{q})} \text{sim}(\boldsymbol{q},\boldsymbol{p}) > \max_{\boldsymbol{n}\in\mathcal{N}(\boldsymbol{q})} \text{sim}(\boldsymbol{q},\boldsymbol{n})\right]. \tag{12}$$

A tie between accurate and inaccurate candidates fails StrictSep. PairAcc measures individual pair orders, while StrictSep measures complete candidate-set separation, making the metrics complementary even at high pairwise accuracy.

As shown in Table 11, all five SLS models exceed the unadapted audio-text models, AF-Next, and Qwen3-Omni on ParaSpeechCaps caption ranking in both metrics. Adapted $PE_{AV}L$ is the strongest caption ranker, while adapted $PE_{AV}S$ has the highest audio ranking PairAcc among SLS variants. For audio ranking, adapted $PE_{AV}S$ ties CLSP and ParaSpeechCLAP at the displayed PairAcc precision. The consistent caption-ranking gains demonstrate SLS's ability to distinguish accurate descriptions from targeted contradictions.

Coco-Nut evaluates Japanese recordings with English consensus captions derived from Japanese listener descriptions, outside the English training setting. All five SLS models improve StrictSep over the corresponding pretrained backbones, and adapted M2D-CLAP has the highest PairAcc among audio-text models. Furthermore, all five SLS adaptations lead the unadapted audio-text models on

Table 12: Ground-truth ranking after removing four accurate candidates from each existing ten-candidate ranking. Caption ranking ranks the original caption against five inaccurate captions for a query recording. Audio ranking ranks the recording originally paired with the query caption against five recordings with contrasting speaking-style tags. Each row highlighted in blue ("+ SLS") adapts the preceding pretrained backbone. **Bold** and underline denote the highest and second/third best scores in each column, respectively.

	Model Parameters		ParaSpeechCaps				Coco-Nut		AudioCaps		MusicCaps	
			Caption		Audio		Caption		Caption		Caption	
Model	Audio	Text	Mean Rank ↓	Acc@1 ↑	Mean Rank ↓	Acc@1 ↑	Mean Rank ↓	Acc@1 ↑	Mean Rank ↓	Acc@1 ↑	Mean Rank ↓	Acc@1 ↑
Random Guess	–	–	3.50	16.7	3.50	16.7	3.50	16.7	3.50	16.7	3.50	16.7
Audio-text Models												
MS-CLAP[23']	0.03B	0.1B	3.05	6.7	3.29	24.6	2.96	21.6	1.31	84.1	2.33	25.2
LAION-CLAP	0.03B	0.1B	3.03	6.6	3.39	21.5	2.97	23.4	1.33	82.7	2.67	19.3
Human-CLAP	0.03B	0.1B	3.14	6.5	3.57	18.1	2.82	19.0	1.06	96.2	2.63	18.7
MGA-CLAP	0.03B	0.1B	2.83	5.6	3.53	17.9	2.70	23.9	**1.01**	99.5	2.54	11.2
GLAP	0.09B	0.8B	3.15	6.2	3.30	24.8	3.54	12.1	1.05	96.5	2.35	26.0
ParaCLAP	0.2B	0.1B	3.35	4.4	3.62	19.4	3.10	20.8	3.39	19.1	3.42	9.7
ParaSpeechCLAP	0.3B	0.3B	1.85	42.1	**1.90**	67.9	3.07	13.3	3.04	25.6	3.32	7.6
CLSP	0.6B	0.1B	2.06	36.3	1.92	67.3	2.39	32.0	2.88	31.4	3.00	12.9
VoiceCLAP L	0.6B	7B	2.42	16.1	2.80	36.9	2.09	41.7	1.04	97.9	2.32	26.2
AuroLA 3B	0.6B	3B	2.50	13.2	2.69	36.1	2.63	23.5	**1.01**	99.3	2.12	31.7
AuroLA 7B	0.6B	7B	2.53	13.3	2.55	40.1	2.39	30.1	**1.01**	99.2	2.04	35.1
OEA Nemotron-3B	0.6B	3B	2.41	26.2	3.13	26.6	2.48	28.2	1.02	98.9	2.58	19.7
OEA Qwen-3B	0.6B	3B	2.58	12.9	3.18	23.2	2.36	30.6	1.02	98.7	2.61	20.3
OEA Qwen-7B	0.6B	7B	2.52	14.1	3.14	25.3	2.43	23.4	1.02	98.7	2.51	22.6
LCO 3B	0.6B	3B	3.08	6.7	3.25	20.5	2.48	30.1	1.03	97.8	2.20	30.3
LCO 7B	0.6B	7B	2.98	7.1	3.18	21.9	2.51	27.2	1.02	98.7	2.17	32.9
VoiceCLAP S	0.09B	0.02B	2.84	10.2	2.98	29.5	2.81	29.1	2.96	29.9	3.22	10.8
+ **SLS**	0.09B	0.02B	1.33	73.1	1.97	62.9	2.75	26.6	1.10	92.8	1.80	48.4
M2D-CLAP[25']	0.09B	0.1B	3.38	2.9	3.27	21.2	2.81	21.5	**1.01**	99.6	2.38	24.4
+ **SLS**	0.09B	0.1B	1.34	71.8	1.98	62.1	2.36	31.5	**1.01**	99.5	1.35	71.1
PE_{AV}S	0.09B	0.3B	2.80	14.5	2.75	34.7	2.38	29.4	1.02	98.2	2.22	30.8
+ **SLS**	0.09B	0.3B	1.31	73.7	**1.90**	64.1	2.53	27.2	**1.01**	99.5	1.33	73.4
PE_{AV}B	0.2B	0.3B	2.88	11.6	2.88	32.3	2.18	36.2	1.02	98.4	2.21	29.9
+ **SLS**	0.2B	0.3B	1.30	73.1	2.02	63.0	2.50	31.8	**1.01**	99.3	1.31	75.4
PE_{AV}L	1B	0.3B	2.82	12.2	2.72	36.6	2.47	28.2	1.02	98.2	2.21	31.1
+ **SLS**	1B	0.3B	1.26	77.2	2.05	61.9	2.74	25.8	**1.01**	99.5	1.29	76.8
Audio LLMs												
AF-Next	0.6B	7B	1.87	50.4	3.29	20.2	1.50	70.8	1.39	88.8	1.84	53.3
Kimi-Audio	0.6B	7B	1.70	57.0	3.00	32.7	1.51	67.3	1.26	86.5	1.79	56.5
Qwen3-Omni	0.6B	30B	1.55	77.3	2.86	38.5	1.13	92.7	1.17	95.9	1.38	82.4
Gemini 3.7 Flash	–	–	**1.01**	**99.5**	2.06	**68.1**	**1.00**	**99.7**	**1.01**	**99.8**	**1.00**	**99.9**

MusicCaps ranking in both metrics. AudioCaps is relatively easy for many models, consistent with AudioCaps' short, event-focused descriptions and direct contrasts.

C.4 Ranking Ground Truth Against Negatives

Protocol. We test whether a model ranks the original caption or the paired recording above five negatives. For audio-to-text ranking, we compare each query recording with the original caption and five inaccurate captions. For text-to-audio ranking, we compare each query caption with its original paired recording and five recordings with contrasting speaking-style tags, selected as described in Appendix A.3. We remove the other four accurate candidates from each existing ranking before calculating the metrics. Audio-LLM results reuse the existing ten-candidate responses, keeping the order of the six remaining candidates in each ranked list.

Metrics. Let $\boldsymbol{q}$ be a query recording or caption, $g(\boldsymbol{q})$ its original paired caption or recording, and $\mathcal{N}(\boldsymbol{q})$ the set of five negatives. When the model assigns similarity scores to the candidates, the ground-truth rank is

$$r(\boldsymbol{q}) = 1 + \sum_{\boldsymbol{n}\in\mathcal{N}(\boldsymbol{q})} \left(\mathbb{1}\left[\mathrm{sim}(\boldsymbol{q},\boldsymbol{n}) > \mathrm{sim}(\boldsymbol{q},g(\boldsymbol{q}))\right] + \frac{1}{2}\mathbb{1}\left[\mathrm{sim}(\boldsymbol{q},\boldsymbol{n}) = \mathrm{sim}(\boldsymbol{q},g(\boldsymbol{q}))\right] \right). \quad (13)$$

Candidates with equal scores receive the average of their occupied positions, so two candidates tied for first each receive rank 1.5. When a model returns a ranked list, $r(\boldsymbol{q})$ is the position of the original caption or paired recording among the six candidates. Mean Rank averages $r(\boldsymbol{q})$ across the queries included in the evaluation, giving each query equal weight, and ranges from 1 (best) to 6 (worst). Top-1 accuracy (Acc@1) is the percentage of those queries for which the ground truth ranks above all five negatives. Ties for first place count as incorrect. Uniform random ordering has an expected Mean Rank of 3.5 and Acc@1 of 16.7%. Mean Rank measures average position, while Acc@1 counts how often the original caption or paired recording is the only first-place candidate.

Results. As shown in Table 12, all five SLS-adapted models lower Mean Rank and increase Acc@1 in both ParaSpeechCaps directions relative to the pretrained models. These gains show that SLS improves ranking of the original caption against descriptions with changed attributes, even when the other accurate descriptions are excluded from scoring. For audio ranking, adapted PE_{AV}S offers the lowest Mean Rank and highest Acc@1 among the SLS models and matches ParaSpeechCLAP's Mean Rank of 1.90 at the reported precision. Additionally, ParaSpeechCLAP and CLSP perform well at selecting the original paired recording, while all five SLS models exceed AF-Next, Kimi-Audio, and Qwen3-Omni in both audio-ranking metrics.

On MusicCaps, all five adaptations lower Mean Rank and increase Acc@1 relative to the pretrained models and outperform every unadapted audio-text model on both metrics. Furthermore, M2D-CLAP improves both metrics on Coco-Nut, showing better ranking of English descriptions of Japanese speech. The results extend the comparison of original captions and inaccurate captions to speech, sound, and music.

C.5 Manually Reviewed Speaking-style Contrasts

Construction. Supporting a focused diagnostic of speaking-style contrasts, we construct a paired speaking-style evaluation from the ParaSpeechCaps holdout set. Candidate pairs contain two recordings from the same speaker with contrasting target tags, allowing at most one additional tag difference. We compare the recordings' speaking-style tags without using tags that describe the recording environment. Each audio receives a short caption describing the target attribute that differs between the two recordings, with parallel wording across the two captions. For instance, two recordings from the same speaker with "calm" and "scared" emotion tags, respectively, are paired with captions "The speaker sounds calm." and "The speaker sounds scared." Manual review retains 260 of 300 candidate pairs, comprising 106 speaking-rate, 47 emotion, 58 vocal-delivery (normal vs. whispered), and 49 animated versus calm/sleepy contrasts.

Metrics. Following the paired evaluation in CompA (Ghosh et al., 2024), we measure audio-to-text, text-to-audio, and joint accuracy. For matched audio-text pairs $(\boldsymbol{a}_1,\boldsymbol{t}_1)$ and $(\boldsymbol{a}_2,\boldsymbol{t}_2)$, the per-sample correctness indicators are

$$\begin{aligned} c_{A\to T} &= \mathbb{1}[\mathrm{sim}(\boldsymbol{a}_1,\boldsymbol{t}_1) > \mathrm{sim}(\boldsymbol{a}_1,\boldsymbol{t}_2) \ \wedge\ \mathrm{sim}(\boldsymbol{a}_2,\boldsymbol{t}_2) > \mathrm{sim}(\boldsymbol{a}_2,\boldsymbol{t}_1)]\,, \\ c_{T\to A} &= \mathbb{1}[\mathrm{sim}(\boldsymbol{a}_1,\boldsymbol{t}_1) > \mathrm{sim}(\boldsymbol{a}_2,\boldsymbol{t}_1) \ \wedge\ \mathrm{sim}(\boldsymbol{a}_2,\boldsymbol{t}_2) > \mathrm{sim}(\boldsymbol{a}_1,\boldsymbol{t}_2)]\,, \\ c_{A\leftrightarrow T} &= c_{A\to T}\, c_{T\to A}. \end{aligned} \quad (14)$$

All required comparisons are strict, so a tie fails the corresponding criterion. We average each indicator over the pairs within each category and report percentages. Independent similarities drawn from a common continuous distribution yield random accuracies of 25% in each direction and approximately 16.7% jointly.

Results. Table 13 shows that all five SLS adaptations improve joint accuracy over the corresponding pretrained models on speaking rate, emotion, and animated versus calm/sleepy speech. For instance, M2D-CLAP speaking-rate joint accuracy rises from 2.8% to 74.5%, and adapted PE_{AV}S improves animated versus calm/sleepy joint accuracy from 34.7% to 75.5%. These gains support improved audio-dependent discrimination of the reviewed style contrasts after SLS adaptation.

Table 13: Speaking-style contrast accuracy on manually reviewed same-speaker pairs from ParaSpeechCaps. Each sample contains two recordings and their two matched captions. $A \rightarrow T$ requires both recordings to prefer their matched captions, and $T \rightarrow A$ requires both captions to prefer their matched recordings. $A \leftrightarrow T$ requires both conditions. Each row highlighted in blue ("+ SLS") adapts the preceding pretrained backbone. **Bold** and underline denote the highest and second/third best scores in each column, respectively.

	Model Parameters		Speaking Rate			Emotion			Vocal Delivery			Animated vs. Calm/Sleepy		
Model	Audio	Text	$A \rightarrow T$	$T \rightarrow A$	$A \leftrightarrow T$	$A \rightarrow T$	$T \rightarrow A$	$A \leftrightarrow T$	$A \rightarrow T$	$T \rightarrow A$	$A \leftrightarrow T$	$A \rightarrow T$	$T \rightarrow A$	$A \leftrightarrow T$
Random Guess	–	–	25.0	25.0	16.7	25.0	25.0	16.7	25.0	25.0	16.7	25.0	25.0	16.7
Audio-text Models														
MS-CLAP[23']	0.03B	0.1B	2.8	8.5	0.9	36.2	53.2	29.8	**100.0**	**100.0**	**100.0**	79.6	73.5	65.3
LAION-CLAP	0.03B	0.1B	30.2	20.8	9.4	8.5	29.8	6.4	84.5	94.8	81.0	69.4	40.8	30.6
Human-CLAP	0.03B	0.1B	26.4	18.9	10.4	23.4	21.3	8.5	62.1	91.4	53.4	75.5	59.2	51.0
MGA-CLAP	0.03B	0.1B	34.9	18.9	9.4	6.4	40.4	4.3	75.9	**100.0**	75.9	34.7	55.1	24.5
GLAP	0.09B	0.8B	27.4	6.6	5.7	21.3	17.0	8.5	74.1	20.7	19.0	14.3	24.5	6.1
ParaCLAP	0.2B	0.1B	11.3	3.8	0.9	57.4	59.6	48.9	20.7	12.1	6.9	28.6	65.3	24.5
ParaSpeechCLAP	0.3B	0.3B	**93.4**	90.6	**84.9**	85.1	89.4	**85.1**	98.3	94.8	94.8	**95.9**	95.9	**93.9**
CLSP	0.6B	0.1B	86.8	89.6	80.2	**87.2**	**97.9**	**85.1**	98.3	98.3	96.6	**95.9**	98.0	**93.9**
VoiceCLAP L	0.6B	7B	33.0	54.7	29.2	63.8	83.0	57.4	98.3	**100.0**	98.3	85.7	93.9	81.6
AuroLA 3B	0.6B	3B	40.6	34.0	23.6	74.5	76.6	63.8	98.3	98.3	98.3	81.6	65.3	57.1
AuroLA 7B	0.6B	7B	41.5	42.5	27.4	76.6	83.0	70.2	**100.0**	**100.0**	**100.0**	85.7	79.6	69.4
OEA Nemotron-3B	0.6B	3B	22.6	26.4	16.0	48.9	61.7	29.8	98.3	**100.0**	98.3	28.6	24.5	12.2
OEA Qwen-3B	0.6B	3B	11.3	18.9	2.8	57.4	72.3	46.8	98.3	**100.0**	98.3	79.6	26.5	18.4
OEA Qwen-7B	0.6B	7B	21.7	17.9	6.6	51.1	78.7	42.6	98.3	**100.0**	98.3	91.8	65.3	61.2
LCO 3B	0.6B	3B	29.2	11.3	6.6	59.6	46.8	44.7	70.7	67.2	58.6	32.7	26.5	14.3
LCO 7B	0.6B	7B	26.4	15.1	10.4	61.7	51.1	44.7	79.3	67.2	58.6	32.7	20.4	12.2
VoiceCLAP S	0.09B	0.02B	28.3	67.9	27.4	46.8	40.4	29.8	79.3	84.5	70.7	59.2	71.4	55.1
+ **SLS**	0.09B	0.02B	70.8	85.8	69.8	80.9	89.4	78.7	**100.0**	**100.0**	**100.0**	93.9	**100.0**	**93.9**
M2D-CLAP[25']	0.09B	0.1B	3.8	31.1	2.8	14.9	25.5	4.3	**100.0**	93.1	93.1	51.0	44.9	22.4
+ **SLS**	0.09B	0.1B	76.4	90.6	74.5	76.6	83.0	72.3	98.3	98.3	98.3	73.5	98.0	73.5
PE_{AV}S	0.09B	0.3B	72.6	85.8	67.9	48.9	48.9	34.0	98.3	96.6	96.6	44.9	65.3	34.7
+ **SLS**	0.09B	0.3B	76.4	85.8	73.6	83.0	89.4	78.7	**100.0**	**100.0**	**100.0**	75.5	98.0	75.5
PE_{AV}B	0.2B	0.3B	68.9	82.1	64.2	59.6	61.7	48.9	**100.0**	98.3	98.3	65.3	63.3	51.0
+ **SLS**	0.2B	0.3B	78.3	87.7	75.5	80.9	87.2	78.7	**100.0**	**100.0**	**100.0**	91.8	98.0	89.8
PE_{AV}L	1B	0.3B	40.6	82.1	37.7	66.0	61.7	55.3	**100.0**	**100.0**	**100.0**	83.7	61.2	53.1
+ **SLS**	1B	0.3B	77.4	**94.3**	76.4	85.1	93.6	**85.1**	**100.0**	**100.0**	**100.0**	73.5	98.0	73.5

ParaSpeechCLAP and CLSP achieve higher speaking-rate joint accuracy than all SLS variants. Adapted PE_{AV}L matches both specialists on emotion, and adapted VoiceCLAP S matches both on animated versus calm/sleepy speech. Vocal delivery offers limited separation among strong models, with several pretrained and adapted models reaching 100% joint accuracy. The relative strengths of SLS-adapted models and specialist baselines vary across attributes.

Table 14: SLS performance across fine-tuning seeds 0, 1, and 2. Average rows report the mean ± sample standard deviation across the three seeds. Retrieval and ranking averages summarize the two ParaSpeechCaps directions, and the classification average gives equal weight to the four speech families. Retrieval, classification, and ranking use Recall@5, macro F1, and PairAcc, respectively. Other experiments use seed 0 unless stated otherwise.

	Retrieval (Recall@5)			Classification (Macro F1)					Ranking (PairAcc)		
Model	Audio→Text	Text→Audio	Average	Speaking Style	Recording Conditions	Conversation	Content	Average	Caption	Audio	Average
VoiceCLAP S	6.4	4.9	5.7	35.8	20.8	34.7	3.4	23.7	70.8	60.1	65.5
+ **SLS** (seed = 0)	55.5	58.4	57.0	49.6	64.3	56.2	79.2	62.3	96.8	79.4	88.1
+ **SLS** (seed = 1)	54.5	57.6	56.1	48.3	60.7	58.6	80.5	62.0	96.7	79.5	88.1
+ **SLS** (seed = 2)	53.5	57.3	55.4	48.6	64.1	57.6	80.7	62.7	96.9	79.0	88.0
+ **SLS** (average)	54.5±1.0	57.8±0.6	56.2±0.8	48.8±0.7	63.0±2.0	57.5±1.2	80.1±0.8	62.3±0.4	96.8±0.1	79.3±0.3	88.1±0.1
M2D-CLAP[25']	3.6	3.3	3.5	33.1	45.9	35.4	3.1	29.4	56.4	55.8	56.1
+ **SLS** (seed = 0)	61.3	62.9	62.1	45.2	67.4	61.3	86.3	65.0	96.5	80.5	88.5
+ **SLS** (seed = 1)	58.2	60.9	59.6	45.3	64.0	60.5	86.2	64.0	96.7	79.2	88.0
+ **SLS** (seed = 2)	62.2	60.9	61.6	46.5	60.4	63.5	87.1	64.4	96.4	79.6	88.0
+ **SLS** (average)	60.6±2.1	61.6±1.2	61.1±1.3	45.7±0.7	63.9±3.5	61.8±1.6	86.5±0.5	64.5±0.5	96.5±0.2	79.8±0.7	88.2±0.3
$PE_{AV}S$	15.5	16.0	15.8	46.8	51.5	44.5	46.5	47.3	65.1	64.8	65.0
+ **SLS** (seed = 0)	73.6	74.0	73.8	52.7	86.2	65.1	91.9	74.0	97.3	81.5	89.4
+ **SLS** (seed = 1)	71.8	74.2	73.0	54.1	81.0	65.4	92.1	73.1	97.4	81.6	89.5
+ **SLS** (seed = 2)	72.0	73.3	72.7	52.7	84.1	64.8	91.5	73.2	97.3	82.0	89.7
+ **SLS** (average)	72.5±1.0	73.8±0.5	73.2±0.6	53.2±0.8	83.8±2.6	65.1±0.3	91.8±0.3	73.4±0.5	97.3±0.1	81.7±0.3	89.5±0.2

C.6 Random Seeds

Table 14 shows consistent SLS gains across seeds 0, 1, and 2 for VoiceCLAP S, M2D-CLAP, and $PE_{AV}S$. Every tested seed improves retrieval, all four speech classification-family averages, and both ranking directions over the corresponding pretrained backbone. Caption ranking varies little across seeds, and $PE_{AV}S$ maintains the strongest retrieval and classification averages among the three adapted backbones. The average rows report means and sample standard deviations across the three seeds. Seed 0 is used by default elsewhere. The small seed variation relative to the adaptation gains supports the repeatability of the complete training recipe.

Table 15: Audio-negative mining and supervision settings for VoiceCLAP S. Anchor and candidate columns specify the representations used to score eligible audio recordings. N_A^- is the maximum number of mined recordings per anchor, and λ weights local SupCon relative to SigLIP in Eq. 5. Retrieval, classification, and ranking use Recall@5, macro F1, and PairAcc. Cells highlighted in blue mark default settings. Repeated default rows provide references for each comparison block.

	Mining Similarity Metric				VoiceCLAP S			
	Anchor	Candidate	N_A^-	λ	Retrieval	Classification	Caption Ranking	Audio Ranking
No SupCon Loss								
(1)	Pretrained				5.7	23.7	70.8	60.1
(2)	SigLIP with Structured Captions				57.4	61.0	96.8	78.9
Text → Text Mining								
(3)	$\mathcal{T}_i^+$	$\mathcal{T}_j^+$	8	1	56.7	63.9	97.1	79.8
(4)	$\mathcal{T}_i^+$	$\mathcal{T}_j^-$	8	1	43.2	62.4	97.7	77.8
(5)	$\mathcal{T}_i^-$	$\mathcal{T}_j^+$	8	1	56.0	64.3	97.1	78.8
(6)	$\mathcal{T}_i^-$	$\mathcal{T}_j^-$	8	1	55.9	63.9	97.0	79.6
Text → Audio Mining								
(7)	$\mathcal{T}_i^+$	$\boldsymbol{a}_j$	8	1	56.5	62.0	96.7	79.8
(8)	$\mathcal{T}_i^-$	$\boldsymbol{a}_j$	8	1	57.0	62.3	96.8	79.4
Audio → Text Mining								
(9)	$\boldsymbol{a}_i$	$\mathcal{T}_j^+$	8	1	53.9	63.3	97.1	79.7
(10)	$\boldsymbol{a}_i$	$\mathcal{T}_j^-$	8	1	20.0	63.5	98.2	61.8
Audio → Audio Mining								
(11)	$\boldsymbol{a}_i$	$\boldsymbol{a}_j$	8	1	55.1	62.2	96.6	78.7
Number of Mined Audio								
(12)	$\mathcal{T}_i^-$	$\boldsymbol{a}_j$	1	1	44.3	62.9	98.3	73.3
(13)	$\mathcal{T}_i^-$	$\boldsymbol{a}_j$	2	1	51.2	63.5	97.7	76.7
(14)	$\mathcal{T}_i^-$	$\boldsymbol{a}_j$	4	1	55.8	63.5	97.1	78.7
(15)	$\mathcal{T}_i^-$	$\boldsymbol{a}_j$	8	1	57.0	62.3	96.8	79.4
(16)	$\mathcal{T}_i^-$	$\boldsymbol{a}_j$	16	1	57.4	63.0	96.6	79.2
SupCon Loss Weight								
(17)	$\mathcal{T}_i^-$	$\boldsymbol{a}_j$	8	0.2	57.4	61.7	96.8	79.5
(18)	$\mathcal{T}_i^-$	$\boldsymbol{a}_j$	8	0.5	56.7	62.1	96.9	79.1
(19)	$\mathcal{T}_i^-$	$\boldsymbol{a}_j$	8	1	57.0	62.3	96.8	79.4
(20)	$\mathcal{T}_i^-$	$\boldsymbol{a}_j$	8	2	58.2	62.8	96.8	79.3

D Ablation Studies

We examine how audio-negative mining, caption composition, batch size, and training duration affect SLS. These studies extend the component comparisons in Section 3.4 and characterize the training recipe under different supervision and compute settings.

D.1 In-batch Audio-negative Mining

Table 15 compares four ways to score eligible in-batch audio candidates. Audio-to-audio (A2A) compares the anchor and candidate recordings. Audio-to-text (A2T) compares the anchor recording with the candidate's captions. Text-to-audio (T2A) compares the anchor's captions with candidate recordings, and text-to-text (T2T) compares their caption sets. These relations identify the representations used to score candidate recordings for mining. In Eq. 15, $\mathcal{X}$ is the anchor audio singleton $\{\boldsymbol{a}_i\}$ or caption set $\mathcal{T}_i^+$ or $\mathcal{T}_i^-$, and $\mathcal{Y}$ is the candidate singleton $\{\boldsymbol{a}_j\}$ or caption set $\mathcal{T}_j^+$ or $\mathcal{T}_j^-$.

$$\mathcal{M}_i = \left\{ \boldsymbol{a}_j : j \in \arg\operatorname*{top\text{-}}_{j \in \mathcal{E}_i} N_A^- \frac{1}{|\mathcal{X}|\,|\mathcal{Y}_j|} \sum_{\boldsymbol{x} \in \mathcal{X}} \sum_{\boldsymbol{y} \in \mathcal{Y}_j} \mathrm{sg}(\boldsymbol{x})^\top \mathrm{sg}(\boldsymbol{y}) \right\}. \tag{15}$$

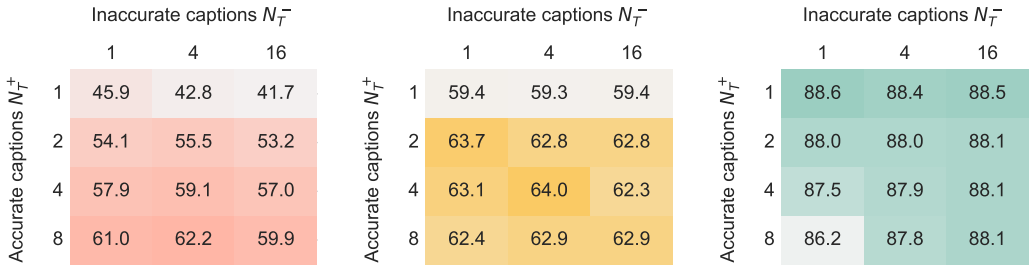
The default SLS adaptation uses $\mathcal{X} = \mathcal{T}_i^-$ and $\mathcal{Y}_j = \{\boldsymbol{a}_j\}$, directly connecting targeted inaccurate captions to audio selection.

Table 16: Example of in-batch audio-negative mining with pretrained $PE_{AV}S$ and SLS-adapted $PE_{AV}S$, using the mining rule in Eq. 1. Candidate recordings are ordered by mean similarity to the anchor's inaccurate captions, with associated reference descriptions shown. Words highlighted in **green** and **red** mark attributes that match and contrast with the anchor description, respectively.

Rank	$PE_{AV}S$	SLS-$PE_{AV}S$
Anchor's ground truth caption: A **male** speaker with a **deep**, **low-pitched** voice delivers **measured** and **flowing** speech in an **American** accent, recorded against a backdrop of a rather **noisy** environment. The silky texture of his voice adds to the clarity of his speech despite the surrounding disturbances.		
1	A **male** speaker with a **low-pitched**, **American** accent delivers **measured**, deliberate speech in a **very noisy** environment, giving his voice a **deep**, **resonant** quality.	A **male** speaker with a **guttural**, **husky** voice and a **measured** speaking speed delivers his words in a **raspy** tone, recorded in a **very noisy American** environment with a **low pitch**.
2	A **German-accented male** speaker from the **middle-aged age** group is talking.	A **male** speaker's **deep**, **low-pitched** voice is heard, speaking at a **measured** speed in a **very noisy American** environment. The volume of his speech is **loud**.
3	A **male** speaker with a **deep** voice and a **French** accent delivers his speech at a **measured** pace in a **noisy** environment. His tone is **hesitant**, and his voice has a **low-pitched**, **deep** quality.	A **male** speaker with a **low-pitched**, **American** accent delivers **measured**, deliberate speech in a **very noisy** environment, giving his voice a **deep**, **resonant** quality.
4	**A person asks to increase the heat in the washroom.**	A **British male** speaks in a **monotonous**, **measured** tone with a **low-pitched** voice, displaying **hesitancy** in his speech. The recording takes place in a **very noisy** environment, making it challenging to discern every word clearly. Despite the background disturbance, the speech maintains a **flowing** rhythm.
5	A **male** speaker's **deep**, **low-pitched** voice is heard, speaking at a **measured** speed in a **very noisy American** environment. The volume of his speech is **loud**.	A **male** speaker with an **American** accent delivers a speech at a **measured** speed, characterized by a **flowing** rhythm. The recording takes place in a **noisy** environment, yet his voice remains prominent due to its **booming**, **medium-pitched** tone.
6	A **man** speaks in authoritative English, delivering his words **quickly** in a **deep**, **low-pitched** voice against a backdrop of a **very noisy** environment. The **flowing** nature of his speech remains clear and expressive despite the **fast pace** and disruptive noise.	A **British male** speaks with a **deep**, **measured** pace in a **very noisy** environment. His voice is **low-pitched** and **flowing**.
7	**This recording has speech from one speaker. The conversation contains exactly one speaking turn. A recording with non-overlapping speech.**	A **male American** speaker's crisp, **medium-pitched** voice flows steadily despite the **very noisy** environment, delivering his words at a **measured** speed.
8	A **male** speaker with an **American** accent delivers authoritative, **fast-paced** speech in a **noisy** environment. His voice is **husky**, **booming**, and **raspy** with a **medium-pitched**, **deep** tone.	A **male** speaker with a **guttural American** accent delivers **deep**, **flowing**, and slightly **slurred** speech in a **slightly clean** environment. His **slow** speed and **low-pitched** voice add to the speech's deliberate and rhythmic quality.

Rows (2) to (11) show strong classification and ranking under several mining choices, with row (2) providing the structured-caption SigLIP control. T2T mining gives particularly strong classification, while the default rule balances strong retrieval with high performance in both ranking directions (row 8). Audio-to-inaccurate-caption mining emphasizes caption ranking, but yields lower retrieval and audio ranking (row 10). In rows (12) to (15), increasing mined recordings from one to eight improves retrieval and audio ranking while maintaining high caption-ranking accuracy. $N_A^- = 16$ offers similar performance, supporting a compact set of informative audio contrasts. Rows (17) to (20) show modest sensitivity across the tested SupCon weights, supporting the recipe's stability across practical supervision settings.

Table 16 illustrates audio-negative mining with pretrained and SLS-adapted $PE_{AV}S$ through the candidates' reference descriptions. In the left part of the table, the pretrained $PE_{AV}S$ already selects useful recordings whose descriptions share anchor attributes, including low pitch and measured delivery. After SLS adaptation, the selected descriptions concentrate on detailed vocal style, combining shared attributes with contrasts in texture, pitch, accent, and delivery. The leading adapted candidate preserves several anchor attributes while contrasting a guttural, husky voice with the anchor's silky quality. The displayed descriptions illustrate how adaptation focuses audio mining on close stylistic alternatives.



(a) Retrieval (b) Classification (c) Ranking

Figure 7: Effect of the limits on accurate captions N_T^+ and inaccurate captions N_T^- for VoiceCLAP S. Retrieval averages Recall@5 over both ParaSpeechCaps directions. Classification averages macro F1 over the four speech families. Ranking averages PairAcc over both audio-text directions. All scores are on a 0 to 100 scale. The default uses $N_T^+ = 4$ and $N_T^- = 16$.

Table 17: Effect of global batch size and training duration on VoiceCLAP S with SLS. Rows (1) to (3) vary batch size at 8,000 steps, and rows (4) to (8) vary training steps at batch size 80. Retrieval, classification, and caption/audio ranking use Recall@5, macro F1, and PairAcc on a 0 to 100 scale. Cells highlighted in blue mark default settings.

			VoiceCLAP S			
	Batch Size B	Training Steps	Retrieval	Classification	Caption Ranking	Audio Ranking
(1)	20	8,000	35.8	51.0	97.4	76.5
(2)	40	8,000	47.1	59.0	97.1	77.9
(3)	80	8,000	57.0	62.3	96.8	79.4
(4)	80	1,000	28.3	53.6	96.7	74.9
(5)	80	2,000	38.4	57.6	97.1	76.5
(6)	80	4,000	48.2	60.7	97.2	78.0
(7)	80	8,000	57.0	62.3	96.8	79.4
(8)	80	16,000	63.6	64.5	96.3	79.1

D.2 Structured Caption Composition

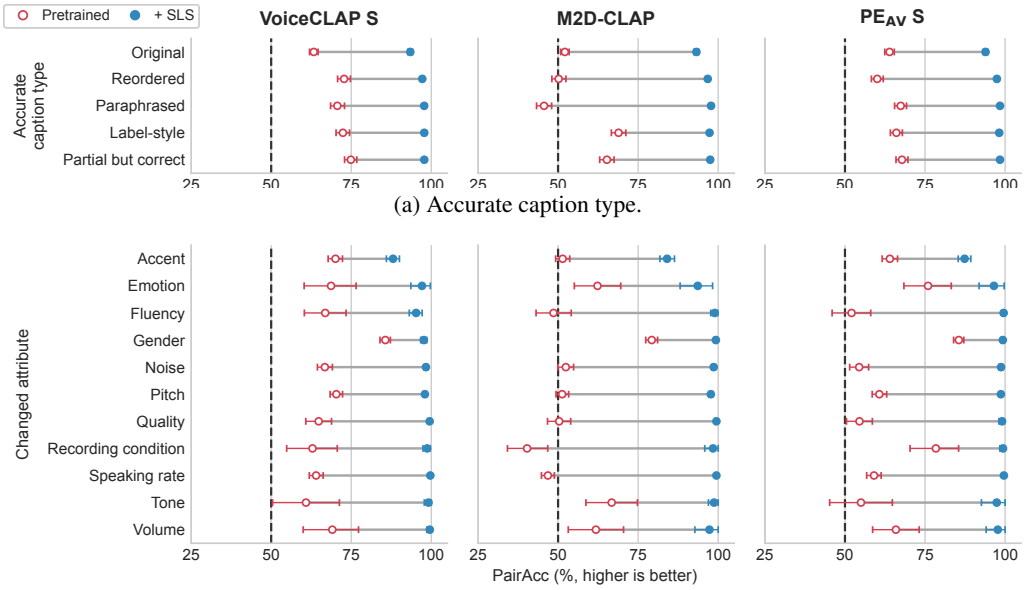
Figure 7 depicts the SLS-adapted VoiceCLAP S performance under different N_T^+ and N_T^-. First, more accurate captions consistently improve retrieval, while classification benefits most from moving beyond one positive. Additional inaccurate captions support ranking when accurate pools are larger. Note that the default was selected using the broader speech, sound, and music validation sets.

D.3 Batch Size Scaling

At 8,000 training steps, increasing batch size from 20 to 40 to 80 improves retrieval, classification, and audio ranking in Table 17, while caption ranking stays high. Larger batches provide more in-batch examples for correspondence learning and audio-negative selection. The gains suggest further scaling potential, but available compute limits the current default batch size to 80.

D.4 Training Steps Scaling

Table 17 shows that retrieval and classification continue improving through 16,000 steps. Caption ranking is already strong at 1,000 steps, while audio ranking peaks at 8,000 steps among the displayed settings. The different trajectories indicate rapid acquisition of local caption discrimination alongside continuing broader-task gains. We retain 8,000 steps as the default, combining strong multi-task performance with the shorter training budget.



(a) Accurate caption type.

(b) Changed attribute in inaccurate captions.

Figure 8: Caption discrimination before and after SLS adaptation on ParaSpeechCaps. (a) compares each accurate caption type against all inaccurate captions. (b) compares each changed-attribute group against all accurate captions. Error bars show pointwise 95% query-bootstrap intervals. Dashed lines indicate chance performance (50%).

E Caption Ranking Study

This section analyzes how pretrained and SLS-adapted audio-text models rank accurate and inaccurate captions. Figure 8 reports caption-ranking accuracy by the type of accurate description and the attribute changed in an inaccurate caption.

Accurate descriptions. Figure 8a shows stronger separation of accurate descriptions from inaccurate ones after SLS adaptation across original, reordered, paraphrased, label-style, and partial captions. The gains show improved ranking of valid descriptions despite differences in wording, format, and descriptive detail.

Attribute contrasts. Figure 8b groups inaccurate captions by the changed attribute. SLS increases PairAcc across all three backbones, consistently improving the capability to distinguish accurate descriptions from targeted attribute edits. The two analyses show improvements across different forms of valid descriptions and different attributes changed in inaccurate captions.

Table 18: Successful ParaSpeechCaps caption ranking with pretrained $PE_{AV}S$ (Original) and SLS-adapted $PE_{AV}S$. Entries show raw audio-text dot-product scores and ranks within the same ten candidates, with rank 1 being the highest.

		Score (Rank)	
Type	Caption	Original	SLS
Paraphrase	A woman from Australia projects her smooth, clear voice at a steady pace, remaining articulate even when background noise is present.	1.566 (5)	4.110 (1)
Partial Positive	A female speaker with a medium-pitched, silky voice delivers her words with measured speed in a noisy environment.	1.019 (10)	3.829 (2)
Label Style	speaker=female; accent=australian; pitch=medium-pitched; quality=silky/crisp; speed=measured; volume=loud; background=noisy.	1.343 (6)	3.749 (3)
Reordering	A loud female speaker with an Australian accent talks at a measured speed in a noisy environment, her voice medium-pitched, silky, and crisp.	1.201 (8)	3.687 (4)
Ground Truth	A female speaker with a medium-pitched, silky, and crisp voice delivers her words with measured speed and enunciation, despite speaking in a noisy environment. Her Australian accent adds distinctiveness to her loud and expressive speech.	1.667 (3)	2.710 (5)
Inaccurate	A female speaker with a **high-pitched**, silky, and crisp voice delivers her words with measured speed and enunciation, despite speaking in a noisy environment. Her Australian accent adds distinctiveness to her loud and expressive speech.	1.313 (7)	0.363 (6)
Inaccurate	A female speaker with a medium-pitched, silky, and crisp voice delivers her words with measured speed and enunciation, despite speaking in a **clean** environment. Her Australian accent adds distinctiveness to her loud and expressive speech.	1.796 (1)	–0.637 (7)
Inaccurate	A female speaker with a medium-pitched, silky, and crisp voice delivers her words with **rapid** speed and enunciation, despite speaking in a noisy environment. Her Australian accent adds distinctiveness to her loud and expressive speech.	1.767 (2)	–0.823 (8)
Inaccurate	A female speaker with a medium-pitched, **rough**, and crisp voice delivers her words with measured speed and enunciation, despite speaking in a noisy environment. Her Australian accent adds distinctiveness to her loud and expressive speech.	1.631 (4)	–3.400 (9)
Inaccurate	A **male** speaker with a medium-pitched, silky, and crisp voice delivers her words with measured speed and enunciation, despite speaking in a noisy environment. Her Australian accent adds distinctiveness to her loud and expressive speech.	1.130 (9)	–5.155 (10)

Table 19: ParaSpeechCaps caption-ranking example with incomplete separation after SLS adaptation. Entries show raw audio-text dot-product scores and ranks for pretrained $PE_{AV}S$ (Original) and SLS-adapted $PE_{AV}S$, with rank 1 being the highest.

		Score (Rank)	
Type	Caption	Original	SLS
Inaccurate	A female speaker with an **American** accent delivers authoritative, enunciated words in a punctuated and measured speed, despite the presence of a very noisy environment. Her voice is characterized by a high-pitched tone.	1.304 (5)	1.599 (1)
Partial Positive	A female speaker with a British accent delivers authoritative, enunciated words in a punctuated and measured speed.	1.675 (1)	1.290 (2)
Paraphrase	A woman from Britain speaks with clear, commanding diction and a high-frequency voice, keeping a steady, rhythmic pace even though there is significant background noise.	1.276 (6)	0.805 (3)
Label-style	speaker=female; accent=british; voice=high-pitched; style=authoritative; rate=measured; noise=very noisy.	1.489 (3)	-0.014 (4)
Reordering	Delivering authoritative, enunciated words, a British female speaker uses a high-pitched tone at a punctuated and measured speed within a very noisy environment.	1.015 (9)	-0.191 (5)
Inaccurate	A female speaker with a British accent delivers authoritative, enunciated words in a punctuated and measured speed, despite the presence of a very noisy environment. Her voice is characterized by a **low-pitched** tone.	1.624 (2)	-2.489 (6)
Ground Truth	A female speaker with a British accent delivers authoritative, enunciated words in a punctuated and measured speed, despite the presence of a very noisy environment. Her voice is characterized by a high-pitched tone.	1.271 (7)	-2.734 (7)
Inaccurate	A female speaker with a British accent delivers authoritative, enunciated words in a punctuated and measured speed, despite the presence of a **clear** environment. Her voice is characterized by a high-pitched tone.	1.227 (8)	-5.243 (8)
Inaccurate	A female speaker with a British accent delivers authoritative, enunciated words in a punctuated and **rapid** speed, despite the presence of a very noisy environment. Her voice is characterized by a high-pitched tone.	1.346 (4)	-5.563 (9)
Inaccurate	A **male** speaker with a British accent delivers authoritative, enunciated words in a punctuated and measured speed, despite the presence of a very noisy environment. Her voice is characterized by a high-pitched tone.	0.729 (10)	-7.266 (10)

F INSTRUCTION-CONDITIONED TTS EVALUATION

F.1 DATA AND TTS MODELS

We select 20 utterances from the 1,000-utterance English subset of InstructTTSEval. Selection requires complete synthesis and judge coverage, reference durations of 5 to 12 seconds, and adult speakers. We balance female/male and American/British tags while maintaining diversity in emotion, personality, texture, and tone. The six audio conditions in Table 20 are reference audio, Gemini 3.1 Flash TTS Preview with and without instructions,[28] Hume Octave v1,[29] Qwen3-TTS VoiceDesign 1.7B (Hu et al., 2026),[30] and MOSS-VoiceGenerator (Huang et al., 2026).[31] Across APS, DSD, and RP, the six audio conditions yield 360 evaluated combinations of utterance, audio condition, and instruction type. Instructed systems receive the transcript and corresponding instruction, while Gemini without instructions receives only the transcript. All listening files are loudness-normalized to -23 LUFS for consistent playback preparation. All automatic evaluators, including audio-text models and audio LLMs, receive the same loudness-normalized audio as human listeners.

F.2 HUMAN EVALUATION AND STATISTICAL ANALYSIS

We recruit 150 English-speaking listeners from the USA and the UK through Prolific at $12 USD per hour and administer the survey with Gorilla.[32][33] After brief training, listeners answer three screening questions, each asking which of two speaking-style descriptions better matches a recording (Figure 9a). Those answering at least two correctly enter the main survey. The others are screened out and compensated with $0.14 USD, following the recommendation from Prolific.

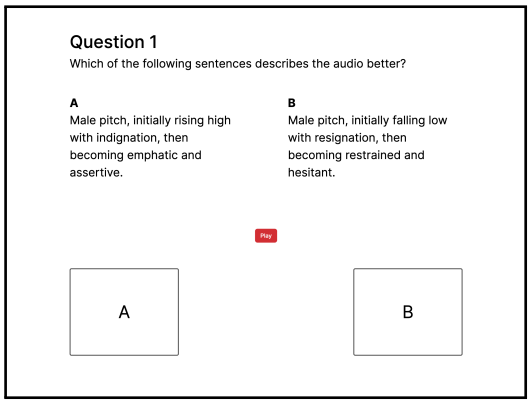
Each listener evaluates one instruction type and two utterances across all six conditions, yielding 12 ratings per participant. APS, DSD, and RP have 49, 50, and 51 listeners, respectively, yielding 1,800 ratings and 3 to 6 ratings per example. All completed submissions were retained.

Attribute-level matching questions support assessment of the 12 explicit APS attributes: gender, pitch, speed, volume, age, clarity, fluency, accent, texture, emotion, tone, and personality. Listeners judge whether each attribute describes the recording (Figures 9b and 9c). DSD and RP examples are shown in Figures 9d and 9e. Note that "volume" in APS reflects attributes like "shouting" and "soft-spoken delivery," which can be perceived even after loudness normalization.

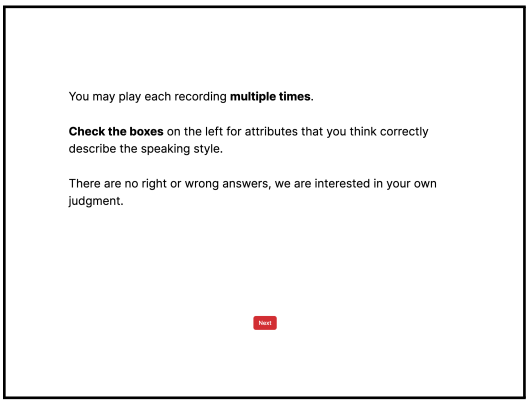
APS scores are the fraction of 12 attributes marked as matching on $[0, 1]$. DSD and RP use five-point adherence ratings similar to the 5-point Likert scale on $[1, 5]$. The Overall score in Table 20 averages instruction types after rescaling DSD and RP to $[0, 1]$.

Statistical analysis. We report participant-level, t-based 95% confidence intervals for human ratings (Table 20), and utterance-clustered bootstrap 95% confidence intervals for correlations (Tables 6 and 21). We assess macro correlation gains over pretrained models using two-sided paired utterance-clustered bootstrap-t tests, with Holm correction across the six adaptation vs. pretrained comparisons separately for each correlation metric.

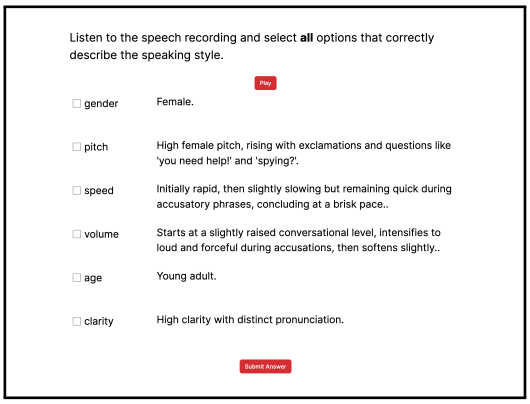
[28] https://deepmind.google/models/model-cards/gemini-3-1-flash-audio/
[29] https://www.hume.ai/octave
[30] https://hf.co/Qwen/Qwen3-TTS-12Hz-1.7B-VoiceDesign
[31] https://hf.co/OpenMOSS-Team/MOSS-VoiceGenerator
[32] https://www.prolific.com
[33] https://gorilla.sc



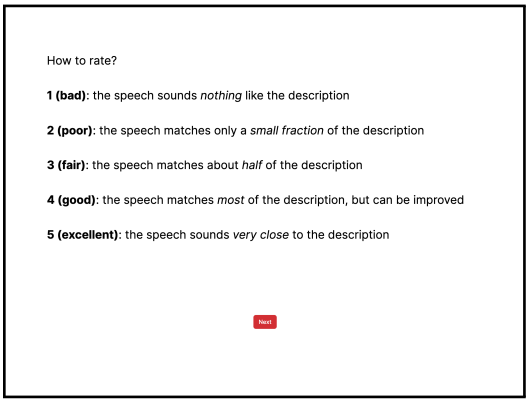
(a) Screening question example.



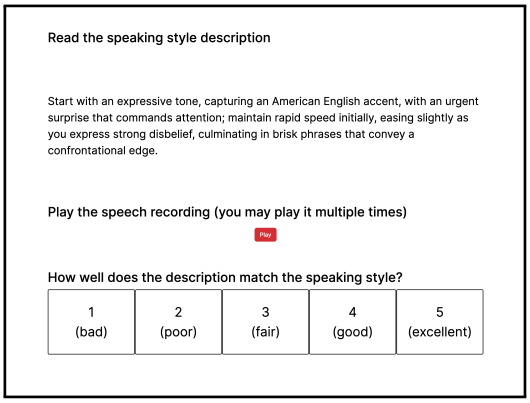
(b) APS-style instruction.



(c) APS-style example.



(d) DSD/RP-style instruction.



(e) DSD/RP-style example.

Figure 9: Example pages of the human evaluation survey. (a) depicts an example two-option question during the screening session. (b) and (c) respectively demonstrate the APS-style instruction and listening test pages. Each APS-style evaluation is split into two pages, so (c) only lists six attributes. (d) and (e) respectively demonstrate the DSD/RP-style instruction and listening test pages.

Table 20: Mean human instruction-adherence ratings and participant-level 95% confidence intervals for six audio conditions on the 20-utterance English InstructTTSEval subset. "Overall" averages across the three instruction types, with DSD and RP rescaled to $[0, 1]$ before aggregation. **Bold** indicates the highest mean.

		InstructTTSEval (English Subset)			
TTS System	Type	APS $[0, 1]$	DSD $[1, 5]$	RP $[1, 5]$	Overall $[0, 1]$
Reference	Real Audio	$0.694_{\pm 0.045}$	$3.55_{\pm 0.25}$	$3.33_{\pm 0.23}$	$0.638_{\pm 0.031}$
Gemini 3.1 Flash TTS Preview	Proprietary	$\mathbf{0.752}_{\pm 0.045}$	$\mathbf{3.66}_{\pm 0.25}$	$\mathbf{3.75}_{\pm 0.25}$	$\mathbf{0.701}_{\pm 0.032}$
Gemini 3.1 Flash TTS Preview (no instruction)	Proprietary	$0.538_{\pm 0.037}$	$2.54_{\pm 0.26}$	$2.54_{\pm 0.31}$	$0.436_{\pm 0.035}$
Hume Octave	Proprietary	$0.532_{\pm 0.045}$	$2.44_{\pm 0.25}$	$2.62_{\pm 0.26}$	$0.432_{\pm 0.033}$
Qwen3-TTS VoiceDesign 1.7B	Open-weight	$0.664_{\pm 0.046}$	$2.90_{\pm 0.21}$	$2.74_{\pm 0.31}$	$0.524_{\pm 0.034}$
MOSS-VoiceGenerator	Open-weight	$0.454_{\pm 0.045}$	$2.58_{\pm 0.27}$	$2.49_{\pm 0.27}$	$0.407_{\pm 0.035}$

F.3 Human Evaluation Results

Table 20 shows that instructed Gemini 3.1 Flash TTS Preview receives the highest mean adherence ratings. The instructed Gemini system exceeds both the reference audio and Gemini without instructions in mean Overall adherence. The Gemini within-model comparison demonstrates the benefit of instruction conditioning, with higher mean adherence ratings across all instruction types. Moreover, Qwen3-TTS VoiceDesign 1.7B is the strongest evaluated open-weight system across all instruction types and Overall.

Notably, the Gemini TTS model's tendency to consistently outperform reference audio with instructions, also reported by Huang et al. (2025), requires careful inspection. InstructTTSEval uses Gemini 2.5 Pro Preview to derive acoustic descriptions from reference recordings as APS instructions, then uses GPT-4o to generate DSD and RP instructions from those descriptions, potentially introducing misalignment between the real audio and text descriptions/instructions. Hence, TTS models may generate speech conditioned on counterfactual instructions that incorrectly describe reference recordings. I.e., correctly following the instructions does not imply synthesizing speech that is close to the reference audio. These reasons offer a possible explanation for why Gemini TTS has higher instruction-adherence scores, as shown in Table 20.

Furthermore, Gemini TTS without instruction scores similarly to Hume Octave. When listening to the actual synthesized audio, we found that Gemini TTS can sometimes infer the emotion from the transcript. For instance, the model synthesizes "oh my gosh" with a more exaggerated tone to express surprise. Nevertheless, instruction-conditioned Gemini TTS and Qwen3-TTS surpass the Gemini TTS without instructions, offering valuable data points for collecting human ratings of TTS instruction adherence.

F.4 Audio-LLM-as-a-Judge

We use AF-Next, Qwen3-Omni, Kimi-Audio, Gemini 3.7 Flash, and Gemini 3.1 Pro Preview to judge TTS instruction adherence. Each model receives speech audio and the corresponding instruction through English prompts (Figures 11 and 10). The prompts ask audio LLMs to judge how closely an audio's overall speaking style matches the description. Similar to human listeners, we ask the audio LLMs to ignore recording quality and transcript accuracy.

For APS (Figure 11), the instruction specifies the same 12 attributes described in Appendix F.2. Judges return one Boolean judgment per attribute and assess each attribute independently. The APS adherence score is the fraction of attributes marked true, yielding a value in $[0, 1]$ consistent with the human-rating definition. For DSD and RP (Figure 10), judges assess overall speaking-style consistency with the provided description and return a single integer score from 1 to 5.

Table 21: Pearson correlations between automatic adherence scores and mean listener ratings for each instruction type. Macro Avg averages APS, DSD, and RP correlations equally. Displayed ± margins give the maximum distance from the estimate to either bound of the 95% utterance-clustered bootstrap interval. ▲ denotes macro gains over the corresponding pretrained model, and ** denotes Holm-adjusted $p < 0.01$. **Bold** indicates the highest correlation in each column and evaluator type.

Automatic Judgment	APS	DSD	RP	Macro Avg
Audio-text Models				
M2D-CLAP	0.154±0.207	0.044±0.177	0.090±0.207	0.096±0.139
+ SigLIP with Structured Captions	0.527±0.122	**0.369**±0.105	0.286±0.171	(▲0.298**) 0.394±0.082
+ **SLS**	0.391±0.138	0.352±0.122	0.300±0.116	(▲0.252**) 0.348±0.081
VoiceCLAP S	**0.597**±0.145	0.095±0.188	**0.396**±0.132	0.363±0.115
+ SigLIP with Structured Captions	0.561±0.117	0.331±0.159	0.270±0.173	(▲0.024) 0.387±0.105
+ **SLS**	0.574±0.128	0.339±0.126	0.306±0.142	(▲0.043) 0.406±0.084
$PE_{AV}S$	0.555±0.110	0.042±0.144	−0.017±0.149	0.193±0.094
+ SigLIP with Structured Captions	0.481±0.129	0.204±0.150	0.352±0.106	(▲0.152) 0.346±0.080
+ **SLS**	0.574±0.102	0.279±0.114	0.385±0.124	(▲0.220**) **0.413**±0.065
ParaSpeechCLAP	0.389±0.116	0.184±0.168	0.257±0.161	0.277±0.087
CLSP	0.574±0.103	0.163±0.201	0.259±0.153	0.332±0.108
Audio LLMs				
AF-Next	0.126±0.126	0.085±0.092	0.181±0.198	0.131±0.085
Qwen3-Omni	0.326±0.210	0.016±0.025	0.164±0.220	0.168±0.103
Kimi-Audio	0.209±0.194	0.147±0.142	0.301±0.164	0.219±0.115
Gemini 3.7 Flash	0.541±0.148	**0.427**±0.173	0.374±0.135	0.447±0.099
Gemini 3.1 Pro Preview	**0.542**±0.147	0.415±0.142	**0.433**±0.123	**0.464**±0.094

```
You are an expert in speech acoustics and expressive speaking style. Listen carefully to the supplied speech
audio and judge how closely its overall speaking style matches the description below.

Consider the audible characteristics requested by the description, including gender presentation, pitch,
speaking speed and rhythm, volume, apparent age, articulation clarity, fluency, accent, vocal texture, emotion,
tone, and personality when they are relevant. First form your own understanding of the audio; do not assume the
description is correct. Honor degree words and penalize clear conflicts. Judge only speaking-style consistency,
ignoring recording quality, speech naturalness, transcript accuracy, and personal preference.

Use this five-point scoring guide:

**1 (bad)**: the speech sounds *nothing* like the description
**2 (poor)**: the speech matches only a *small fraction* of the description
**3 (fair)**: the speech matches about *half* of the description
**4 (good)**: the speech matches *most* of the description, but can be improved
**5 (excellent)**: the speech sounds *very close* to the description

Return exactly one JSON object in the form below. The score must be a JSON integer from 1 through 5. Do not add
explanations, nesting, Markdown, or extra keys.

```json
{
  "score": 5
}
```

Speech style description:

<Insert the speech style description here>
```

Figure 10: Audio-LLM prompt for DSD and RP instruction-adherence scoring. Judges return a single integer rating from 1 to 5 for overall speaking-style consistency with the instruction.

```
You are an expert in speech acoustics and expressive speaking style. Listen carefully to the supplied speech
audio and compare it with the twelve labeled attribute descriptions below.

Judge every attribute independently. For each attribute, output true when the audible speech matches that
attribute's description and false when it does not. A mismatch for one attribute must not influence any other
attribute's judgment.

Use these attribute meanings:

- gender: Vocal characteristics associated with gender presentation, including vocal-fold differences and
socially learned speech patterns.
- pitch: The perceived highness or lowness of the voice, interpreted relative to the described gender
presentation when appropriate.
- speed: How fast or slow the person speaks, including any described changes or rhythmic pattern.
- volume: How loud or soft the speech is, including whispering, conversational speech, shouting, and changes in
loudness.
- age: The speaker's apparent age group or life stage as inferred from speech characteristics.
- clarity: Whether articulation is clear and precise or unclear, slurred, or mumbled.
- fluency: The smoothness and continuity of speech, including hesitation, repetition, and filler words.
- accent: Pronunciation patterns associated with geographic, social, or non-native language background.
- texture: Vocal timbre and texture, such as sweet, hoarse, deep, bright, warm, nasal, soft, rough, or thin.
- emotion: The emotion expressed in the speech, including any change in emotion over time.
- tone: The emotional or attitudinal quality conveyed through pitch movement, such as sarcasm, formality,
enthusiasm, or indifference.
- personality: Personality traits clearly and consistently conveyed by the voice, such as confident, assertive,
anxious, extroverted, or introverted.

Important guidance:

- First form your own understanding of the audio. Do not assume the supplied descriptions are correct.
- Pay close attention to objective conflicts in gender, age, accent, and other salient characteristics.
- Honor degree words. For example, a mildly excited voice does not match a description requiring intense
excitement.
- Judge only the described speaking-style attributes. Ignore recording quality, speech naturalness, transcript
accuracy, and personal preference.
- Produce all twelve judgments, even when one or more descriptions are uncertain.

Return exactly one JSON object with the twelve lowercase keys shown below. Use JSON booleans and do not add
explanations, nesting, Markdown, or extra keys.

```json
{
  "gender": true,
  "pitch": true,
  "speed": true,
  "volume": true,
  "age": true,
  "clarity": true,
  "fluency": true,
  "accent": true,
  "texture": true,
  "emotion": true,
  "tone": true,
  "personality": true
}
```

Twelve-attribute description:

<Insert the twelve-attribute speech style description here>
```

Figure 11: Audio-LLM prompt for APS instruction-adherence scoring. Judges return 12 independent attribute-level Boolean judgments.

Table 22: Comparison of caption relations, supervision, candidate selection, and evaluation scope. SLS combines explicit accurate and inaccurate captions with inaccurate-caption-conditioned audio mining and bidirectional local supervision across existing backbones. Positive grouping, training-time mining, and inference-time selection serve distinct roles. SLS mines audio during training and uses tag-defined candidates for audio-ranking evaluation.

Approach	Relation construction	Supervision	Candidate mining / selection	Evaluation scope
CLSP (Yang et al., 2026)	global and fine-grained positive captions	multi-positive contrastive learning	scheduled positive-caption pairs	speech retrieval, attribute classification, and speech-text similarity against human ratings
SPAM (Cho et al., 2026)	style-key positives	bidirectional multi-positive SupCon; acoustic-attribute regression	positives grouped by shared style key	TTS style adherence, prompt ordering, human alignment
ParaSpeechCLAP (Diwan et al., 2026)	intrinsic/situational captions and tags	bidirectional InfoNCE; variant-specific classification	paired-caption batches	speech retrieval, attributes, TTS selection (best-of-N)
SSE (Zhang et al., 2025b)	generated style descriptions and paralinguistic-label similarities	KL alignment to attribute-weighted soft targets	label-derived within-batch relations	label-matched speech–text retrieval; intra- and inter-corpus evaluation
AuroLA (Xu et al., 2026)	multi-granular captions and tag positives	contrastive learning with hard-negative weighting	tag-based positive grouping	audio-text retrieval
Counterfactual audio (Vosoughi et al., 2024)	generated factual/counterfactual caption pairs	contrastive and counterfactual alignment; angular triplet loss	caption edits supply difficult alternatives	audio concepts and caption discrimination
CompA (Ghosh et al., 2024)	composition-aware caption edits and constructed audio	contrastive and modular composition losses	composition-aware hard negatives	event composition and attribute binding
EmotionRank CLAP (Chandra et al., 2025)	valence-arousal labels	Rank-N-Contrast	label-derived relative orderings	graded emotion distinctions
RA-CLAP (Sun et al., 2025)	teacher similarities	soft relation distillation	teacher-derived batch relations	emotional speaking-style retrieval
RankCLIP (Zhang et al., 2025a)	batch similarities	listwise consistency	within-/cross-modal batch rankings	image-text representation evaluation
Omni-Embed-Audio (Yoo et al., 2026)	user-intent queries and exclusion-based negatives	retrieval adaptation	reformulated and negative queries at evaluation	audio retrieval beyond caption-style queries
SLS	accurate captions and targeted contradictions	SigLIP + bidirectional SupCon	caption-conditioned in-batch audio negatives	retrieval, classification (speech), ranking; TTS adherence transfer

G Extended Related Work

G.1 Closest Training Paradigms

Table 22 compares caption relations, supervision, candidate selection, and evaluation scope across the closest training paradigms, including CLSP and AuroLA. CLSP combines global and fine-grained positive captions through multi-positive speech-text learning (Yang et al., 2026). SPAM uses style-defined positives and bidirectional supervised contrastive learning (Cho et al., 2026), while ParaSpeechCLAP learns rich stylistic speech-text correspondence (Diwan et al., 2026). AuroLA combines multi-granular captions with tag-defined positive grouping and hard-negative weighting (Xu et al., 2026).

Counterfactual audio learning and CompA establish the value of edited descriptions for acoustic and compositional distinctions (Vosoughi et al., 2024; Ghosh et al., 2024). RA-CLAP uses teacher-derived relations (Sun et al., 2025), while EmotionRankCLAP uses valence-arousal orderings (Chandra et al., 2025). SSE uses attribute-weighted paralinguistic label similarities as soft alignment targets (Zhang et al., 2025b).

BACL uses a learned curriculum sampler and token-level local attention to address subtle mismatches (Ye et al., 2025). LCO provides complementary context through language-centric contrastive refinement of generative multimodal representations (Xiao et al., 2025). RankCLIP promotes consistency among batch-level rankings in image-text learning (Zhang et al., 2025a). SLS adds explicit accurate and inaccurate relations, including valid partial descriptions, and uses inaccu-

rate captions to select audio negatives. SPAM, RA-CLAP, SSE, and CompA serve as methodological references because model checkpoints were unavailable during our evaluation.

G.2 Audio Understanding Evaluation and Benchmarking

ParaCLAP, ParaSpeechCLAP, and Vox-Profile emphasize paralinguistic representations and speaker traits (Jing et al., 2024; Diwan et al., 2026; Feng et al., 2025), while FMSU-Bench covers multiple dimensions of speech understanding (Li et al., 2026a). We evaluate SLS on four speech classification families and ranking tasks. Moreover, we evaluate matching on manually reviewed recording pairs from the same speaker with contrasting styles. These tasks test whether correct descriptions and recordings rank above alternatives with contrasting attributes.

Complementary audio scope. T-CLAP's event-order tests, MGA-CLAP's temporal alignment, and FineLAP's clip- and frame-level supervision complement SLS's caption-defined relations (Yuan et al., 2024; Li et al., 2024; 2026c). OEA broadens retrieval evaluation through user-intent reformulations and exclusion-based queries (Yoo et al., 2026). These evaluation queries serve a different purpose from SLS's training-time audio mining.

G.3 Audio Large Language Models

Recent instruction-following benchmarks include InstructTTSEval, MINT-Bench, AnyAudio-Judge, and VStyle (Huang et al., 2025; Chen et al., 2026b; Li et al., 2026b; Zhan et al., 2026). ParaPairAudioBench evaluates audio-LLM judges with pairwise, human-referenced paralinguistic decisions (Jeon et al., 2026). JASTIN and SpeechLLM-as-Judges provide learned evaluator baselines with instruction-conditioned or explanation-based outputs (Zhang et al., 2026; Wang et al., 2026). Our evaluation places audio-text models and audio LLMs under common classification and ranking tasks and compares their adherence scores with listener judgments.

Differences and use cases of audio-text models vs. audio LLMs. Audio-text models encode recordings and descriptions into reusable embeddings, supporting retrieval, classification, and ranking through similarity scores and precomputed candidate representations. Audio LLMs provide a generative interface for instruction following and open-ended responses. Comparisons across these interfaces establish the value of compact SLS-adapted encoders for matching and discrimination tasks alongside the broader capabilities of audio LLMs.

G.4 Instruction-conditioned TTS Evaluation

Lin et al. (2026) study human perception of instruction-guided expressive synthesized speech, motivating perceptually grounded evaluation of expressive control. InstructTTSEval evaluates synthesized speech and validates a Gemini judge using matched and randomly mismatched reference-audio instructions with binary human judgments (Huang et al., 2025). Our study evaluates graded adherence on synthesized outputs and reference speech, comparing audio-text models and audio-LLM judges against listener means. SPAM's instruction-based supervision and Human-CLAP's human relevance judgments offer complementary learned evaluation approaches (Cho et al., 2026; Takano et al., 2025).