QUERY-ORIENTED AND REASONING-AWARE PERCEPTION FOR VERSATILE SPEECH UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing speech understanding systems rely on either implicit speech perception or query-agnostic speech perception, lacking the consideration of the relationship between perception and reasoning. What to perceive depends on the reasoning, while reasoning builds upon perceived information. To address this issue, we propose Query-oriented and Reasoning-aware Perception Optimization (QRPO), a reinforcement learning algorithm that trains a Perceiver to first determine what should be perceived for answering the query and then generate a reasoning-aware description for the perceived information. A Reasoner, which can be instantiated with arbitrary off-the-shelf large language model, performs downstream reasoning based on the description to generate the final response. The correctness of the response is further used as a reward signal to encourage the Perceiver to learn how to describe speech in a way that better supports downstream reasoning. We name the resulting PErceiver-REasoner framework PERE. Experimental results on the MMAU, MMSU, MMAR, and AIR-Bench benchmarks demonstrate that PERE achieves significant gains on speech understanding performance. Furthermore, PERE achieves state-of-the-art results under a more realistic open-ended setting where only the speech and query are provided without candidate answer options.

1 INTRODUCTION

Driven by the rapid advancement of large language models (LLMs) (Brown et al., 2020; OpenAI, 2024; Yang et al., 2025), speech language models (SLMs) have achieved remarkable progress in recent years. Specifically, a growing body of SLMs (Chu et al., 2024; Yu et al., 2026; Gong et al., 2024; Ghosh et al., 2024; 2026) combines audio encoders with powerful language backbones, enabling a unified model to perform a wide range of tasks, including speech recognition, audio captioning, and spoken question answering. More recently, the emergence of reasoning-oriented LLMs (Guo et al., 2025; Shao et al., 2024b) has inspired a new wave of research on speech reasoning (Xie et al., 2025; Wu et al., 2026; He et al., 2026b) and speech agents (Arora et al., 2025; Liang et al., 2026), which seek to integrate speech perception and reasoning to tackle increasingly complex tasks.

Despite these advances, the way existing SLMs handle speech perception remains fundamentally limited. As illustrated in Figure 1(a), most existing SLMs perform implicit speech perception and directly generate the final response (Chu et al., 2024; Wu et al., 2025; Chen et al., 2026). While efficient, the entangled perception and reasoning process makes it difficult to inspect what information has been perceived and how it contributes to the final response. An alternative line of research, shown in Figure 1(b), introduces a separate perception module to explicitly generate descriptions of the speech signal (Tian et al., 2026b; Deshmukh et al., 2025; Wang et al., 2026e). However, this perception process is query-agnostic, generating generic descriptions without considering what information is required for the current query, and thus often failing to capture the evidence most relevant to answering it. More fundamentally, speech perception and reasoning should not be treated as independent processes. Rather, they are intrinsically coupled: what should be perceived is determined by the information required for reasoning, while the effectiveness of reasoning ultimately depends on the quality and relevance of the perceived information (Gou et al., 2025; Bai et al., 2026).

To address these limitations, we propose a modular PErceiver-REasoner framework called PERE. As illustrated in Figure 1(c), a Perceiver is responsible for determining what information should be perceived for answering the given query and how the perceived information should be described to

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.

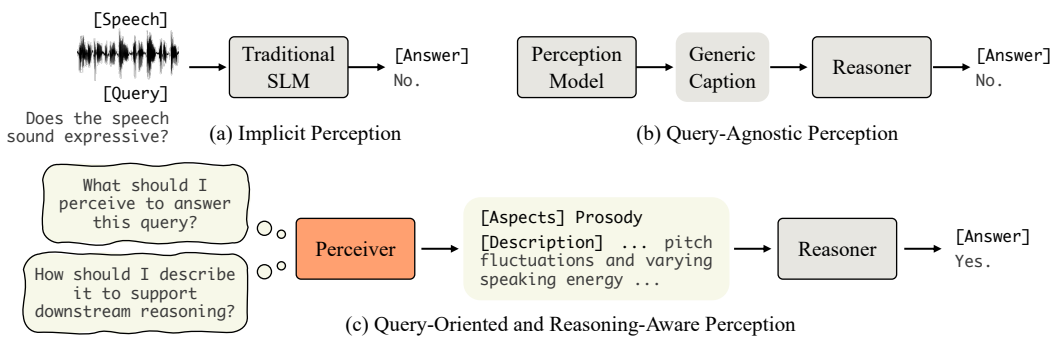

Figure 1: Comparison of speech perception paradigms. (a) Implicit perception in direct-response SLMs. (b) Explicit perception via query-agnostic perception modules. (c) Query-oriented and reasoning-aware perception in the proposed PERE. We omit the inputs in (b) and (c) for simplicity.

best support downstream reasoning. A Reasoner, which can be instantiated with arbitrary off-the-shelf LLM, then performs reasoning over the perceived information to generate the final response. To train the Perceiver, we introduce Query-Oriented and Reasoning-Aware Perception Optimization (QRPO), a reinforcement learning algorithm that aligns speech perception with the requirements of downstream reasoning. QRPO incorporates an aspect selection reward that encourages the Perceiver to dynamically select query-relevant perception aspects, as well as a response reward that uses the correctness of the final answer as reward signals for perception and description quality optimization. As a result, the Perceiver learns not only to identify the information most relevant to a query, but also to present it in a manner that maximally facilitates downstream reasoning.

Beyond the proposed PERE framework, we also revisit the evaluation paradigm for speech understanding. Most existing benchmarks adopt a multiple-choice question answering (MCQ) setting, where candidate answers are provided alongside the speech signal and the query (Sakshi et al., 2025; Wang et al., 2026b; Ma et al., 2026). While convenient for calculating accuracy, candidate options often introduce semantic cues that enable models to partially rely on option matching rather than genuine speech perception and reasoning. Moreover, such a setting departs from real-world scenarios, where candidate answers are typically unavailable. To address this limitation, we complement conventional MCQ evaluation with an open-ended setting, in which models receive only the speech signal and the query, without access to candidate options. This setting places greater demands on both perception and reasoning, providing a more realistic and challenging evaluation protocol.

Our main contributions in this work are fourfold: (1) we propose PERE, a Perceiver-Reasoner framework that explicitly decouples perception from reasoning, with a plug-and-play reasoner that can be arbitrary off-the-shelf LLM; (2) we propose QRPO, a reinforcement learning algorithm that optimizes the Perceiver via a query-oriented aspect selection reward and a reasoning-aware response reward; (3) across the MMAU, MMSU, MMAR, and AIR-Bench benchmarks, PERE obtains significant gains in speech understanding performance; (4) we introduce a more realistic open-ended evaluation setting and show that PERE achieves state-of-the-art results under this setting.

2 Related Work

2.1 Implicit Speech Perception

To leverage the powerful understanding and reasoning abilities of LLMs, recent SLMs are typically constructed by augmenting pretrained LLMs with speech perception capabilities (Chu et al., 2024; Gong et al., 2024; Xie et al., 2025; Tian et al., 2025; Yu et al., 2026). For example, UALM (Tian et al., 2026a) and DualSpeechLM (Wang et al., 2026d) extend the vocabulary of the underlying LLM with speech-specific tokens to enable LLMs to perceive speech. SALMONN-Omni (Yu et al., 2026) and Step-Audio 2 (Wu et al., 2025) project continuous speech representations into the LLM embedding space to achieve speech-text alignment. To improve the reasoning capabilities, research like Audio-CoT (Ma et al., 2025), Audio-Reasoner (Xie et al., 2025), and Audio Flamingo 3 (Ghosh et al., 2026) adopt Chain-of-Thought (CoT) prompting (Wei et al., 2022) to require SLMs to ex-

plicitly generate intermediate thinking process before producing the final response. However, these SLMs perform speech perception in an implicit manner, making it difficult to interpret the information they capture and to systematically optimize the perception process. In addition, due to the modality gap between speech and text (Wang et al., 2026a; Xiang et al., 2025), adapting LLMs to process speech frequently compromises the reasoning capabilities inherited from the backbones.

2.2 Explicit Speech Perception

Explicit speech perception typically produces comprehensive natural language descriptions of the speech signal before determining the final response. For example, EmotionThinker (Wang et al., 2026c) generates prosody-aware descriptions of the speech signal to support emotion reasoning. Bagpiper (Tian et al., 2026b) interprets audio clips through rich captions that capture essential semantic content to facilitate downstream understanding. Similarly, Audio-Mind (Wang et al., 2026e), Caption-before-reason (Deshmukh et al., 2025), and EvA (Xie et al., 2026) introduce a dedicated branch for acoustic evidence extraction before performing deductive reasoning. However, existing approaches to explicit speech perception are largely generic: they generate descriptions without awareness of the given query and are not explicitly optimized to support the subsequent reasoning. In contrast, recent studies in the vision and language domains have increasingly emphasized that perception should be conditioned on the downstream task rather than performed in a task-agnostic manner (Wu & Xie, 2024; Shao et al., 2024a; Zheng et al., 2025; Gou et al., 2025; Bai et al., 2026). Nevertheless, this paradigm remains largely unexplored in the speech domain. To the best of our knowledge, our proposed PERE is the first work to bridge this gap.

3 Methods

As illustrated in Figure 2, PERE is a modular perceiver–reasoner framework comprising a trainable Perceiver and a frozen Reasoner. The key idea behind PERE is to train the Perceiver to perform query-oriented and reasoning-aware perception, enabling it to extract and describe speech evidence that is most relevant to the given query and downstream reasoning. To achieve this, the Perceiver is optimized through a two-stage training process consisting of supervised fine-tuning (SFT) followed by reinforcement learning (RL). In the following sections, we first introduce the training data curation pipeline and then describe the two-stage optimization procedure in detail.

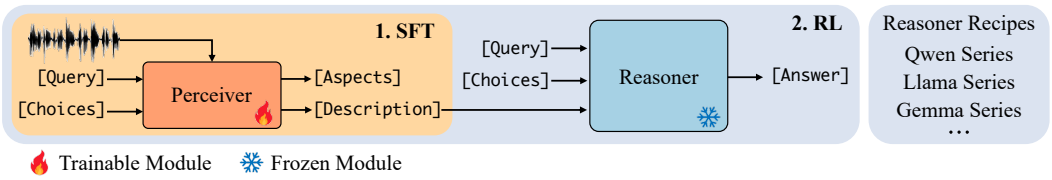

Figure 2: Overview of the PERE framework. Given an input speech, a user query, and optionally a set of candidate choices, the Perceiver first identifies the speech aspects most relevant to answering the query and then generates a reasoning-aware description that contains the corresponding evidence from the speech signal. The Reasoner subsequently takes the query, the optional candidate choices, and the description generated by the Perceiver as input, and performs reasoning to produce the final answer. Because the Reasoner operates exclusively on textual inputs, it can be instantiated with arbitrary off-the-shelf LLM and remains completely frozen throughout the framework.

3.1 Data Curation

Our training data are derived entirely from AudioMCQ (He et al., 2026a), a large-scale audio multiple-choice question answering dataset. Since our focus is on speech understanding rather than general audio or music understanding, we restrict our data to the SpeechCraft (Jin et al., 2024) subset. Formally, let the source corpus be $\mathcal{D}_{\text{src}} = \{(a_i, q_i, O_i, y_i, R_i)\}_{i=1}^{N}$, where N is the number of samples in the source corpus, a_i denotes the speech signal, q_i the query, $O_i = \{c_{i,1}, c_{i,2}, c_{i,3}, c_{i,4}\}$ the candidate choices, $y_i \in \{1, 2, 3, 4\}$ the ground-truth answer index, and R_i the CoT annotation. As illustrated in Figure 3, we augment each speech sample with a query-oriented aspect set and a reasoning-aware description to construct the training corpus $\mathcal{D}_{\text{PERE}}$.

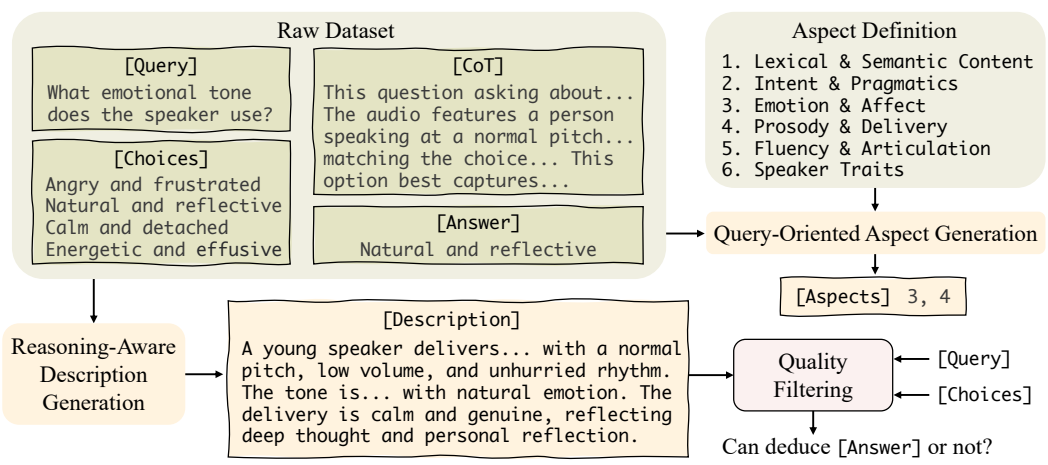

Figure 3: Overview of the training data curation pipeline.

3.1.1 Query-Oriented Aspect Generation

We define a taxonomy of six perception aspects that cover the major perceptual dimensions of speech: $\mathcal{A} = \{\alpha^{\text{lex}}, \alpha^{\text{int}}, \alpha^{\text{emo}}, \alpha^{\text{pro}}, \alpha^{\text{flu}}, \alpha^{\text{spk}}\}$, corresponding to (1) *Lexical & Semantic Content*, (2) *Intent & Pragmatics*, (3) *Emotion & Affect*, (4) *Prosody & Delivery*, (5) *Fluency & Articulation*, and (6) *Speaker Traits*. This taxonomy is grounded in established lines of speech processing research, which characterize the information conveyed by speech in terms of linguistic content, semantics, paralinguistic cues, and speaker-specific attributes (Yang et al., 2021; Schuller & Batliner, 2013; Wang et al., 2026b). For each sample $(a_i, q_i, O_i, y_i, R_i) \in \mathcal{D}_{\text{src}}$, we apply an aspect selection function ϕ implemented with Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025):

$$\phi(q_i, O_i, R_i, y_i) = \mathcal{A}_i \subseteq \mathcal{A}, \tag{1}$$

where $\mathcal{A}_i$ denotes the minimal subset of perception aspects that the Perceiver needs to consider in order to answer the query q_i. Detailed definitions of the six perception aspects and the prompts used for the aspect selection function are provided in Appendix A and Appendix B.1, respectively.

3.1.2 Reasoning-Aware Description Generation

In PERE, reasoning is delegated entirely to the Reasoner, as modern LLMs generally exhibit substantially stronger reasoning capabilities than SLMs (Guo et al., 2025; Wang et al., 2026a; Xiang et al., 2025). Consequently, the role of the Perceiver is to faithfully perceive the speech signal along query-oriented aspects and provide the Reasoner with the description necessary for downstream reasoning. However, the original CoTs in AudioMCQ contain both perceptual observations and reasoning steps leading to the final answer, making them unsuitable for directly supervising the Perceiver. We therefore define a description extraction function ψ, also implemented with Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025), which rewrites the perceptual content in R_i into a concise, reasoning-aware description d_i while removing explicit reasoning traces and answer derivations:

$$\psi(q_i, O_i, R_i, y_i, \mathcal{A}_i) = d_i. \tag{2}$$

To ensure that each description contains sufficient evidence for downstream reasoning, we apply a reasoning-sufficiency filter $\mathcal{M}_{\text{filter}}$ implemented with Qwen2-7B-Instruct (Yang et al., 2024a). Let $\hat{y}_i = \mathcal{M}_{\text{filter}}(q_i, O_i, d_i)$ denote the answer inferred by $\mathcal{M}_{\text{filter}}$ from the query, candidate choices, and generated description alone. The filter retains a sample if and only if the inferred answer matches the ground-truth answer, *i.e.*, $\mathbb{I}[\hat{y}_i = y_i] = 1$. The resulting training corpus $\mathcal{D}_{\text{PERE}}$ is defined as

$$\mathcal{D}_{\text{PERE}} = \{(a_i, q_i, O_i, y_i, \mathcal{A}_i, d_i) : \hat{y}_i = y_i\}, \tag{3}$$

where $|\mathcal{D}_{\text{PERE}}| = M \leq N$ denotes the number of samples after filtering. Following He et al. (2026a), we further partition $\mathcal{D}_{\text{PERE}}$ into a weak audio-contribution subset, $\mathcal{D}^{\text{weak}}_{\text{PERE}}$, and a strong audio-contribution subset, $\mathcal{D}^{\text{strong}}_{\text{PERE}}$. The former subset is used for SFT stage, while the latter subset is used for RL stage. The prompts used for description extraction function and filtering are provided in Appendix B.2 and Appendix B.3, respectively.

3.2 SUPERVISED FINE-TUNING

The Perceiver in PERE is initialized from a direct-response SLM. The objective of supervised fine-tuning is to transform this backbone into a dedicated perception module that first identifies the perception aspects relevant to the query and then faithfully generates a structured description, rather than directly producing an answer. However, naively eliminating the answer-generation behavior may inadvertently degrade the model's speech perception capabilities (see Section 5.2). To mitigate this issue, we duplicate $\mathcal{D}_{\text{PERE}}^{\text{weak}}$ and optimize the Perceiver using two SFT objectives. The same speech-query-choice tuple (a_i, q_i, O_i) is wrapped with different task prompts: p_{perc} for perception-only SFT and p_{ans} for answer-preserving SFT. The p_{perc} and p_{ans} prompts are provided in Appendix B.4 and Appendix B.5, respectively.

Perception-only SFT. Under prompt p_{perc}, for each sample $(a_i, q_i, O_i, y_i, \mathcal{A}_i, d_i) \in \mathcal{D}_{\text{PERE}}^{\text{weak}}$, the Perceiver is trained to generate a structured target response t_i that explicitly encodes the ground-truth aspect set $\mathcal{A}_i$ and the corresponding description d_i in a predefined format:

$$t_i = \texttt{<aspects>}\, \mathcal{A}_i \,\texttt{</aspects><description>}\, d_i \,\texttt{</description>}. \tag{4}$$

Answer-preserving SFT. Under prompt p_{ans}, for the duplicated copies, the Perceiver is instructed to produce the final answer from the same speech, query, and candidate choices. Let c_{i,y_i} denote the ground-truth answer text corresponding to answer index y_i. The model is supervised to generate a structured target response u_i in the form:

$$u_i = \texttt{<answer>}\, c_{i,y_i} \,\texttt{</answer>}. \tag{5}$$

Let π_θ denote the Perceiver policy parameterized by θ. The two SFT objectives are optimized with standard token-level cross-entropy losses:

$$\mathcal{L}_{\text{SFT}}^{\text{perc}}(\theta) = -\mathbb{E}_{(a_i,q_i,O_i,\mathcal{A}_i,d_i)\sim\mathcal{D}_{\text{PERE}}^{\text{weak}}} \left[\log \pi_\theta(t_i \mid a_i, q_i, O_i, p_{\text{perc}})\right], \tag{6}$$

$$\mathcal{L}_{\text{SFT}}^{\text{ans}}(\theta) = -\mathbb{E}_{(a_i,q_i,O_i,y_i)\sim\mathcal{D}_{\text{PERE}}^{\text{weak}}} \left[\log \pi_\theta(u_i \mid a_i, q_i, O_i, p_{\text{ans}})\right], \tag{7}$$

and the overall SFT objective is defined as

$$\mathcal{L}_{\text{SFT}}(\theta) = \mathcal{L}_{\text{SFT}}^{\text{perc}}(\theta) + \mathcal{L}_{\text{SFT}}^{\text{ans}}(\theta). \tag{8}$$

This dual-objective SFT procedure equips the Perceiver with the basic perception behavior and output format. In the subsequent RL training, we start from the SFT checkpoint and further optimize the Perceiver to perform query-oriented and reasoning-aware perception.

3.3 QUERY-ORIENTED AND REASONING-AWARE PERCEPTION OPTIMIZATION

Inspired by Group Relative Policy Optimization (GRPO) (Shao et al., 2024b), originally developed for enhancing LLM performance on mathematical reasoning tasks, we propose Query-Oriented and Reasoning-Aware Perception Optimization (QRPO) for training Perceivers in speech understanding. QRPO encourages the Perceiver to dynamically select appropriate perception aspects conditioned on the current query. Moreover, it explicitly bridges perception and reasoning by using the Reasoner's output as a reward signal to optimize the perception and description quality. Consequently, the perception process of the Perceiver and the reasoning process of the Reasoner are well-aligned.

The overview of QRPO is illustrated in Figure 4. For a training sample $(a_i, q_i, O_i, y_i, \mathcal{A}_i, d_i) \in \mathcal{D}_{\text{PERE}}^{\text{strong}}$, the Perceiver policy π_θ samples a group of G serialized outputs $\{o_i^g\}_{g=1}^G$, where each o_i^g is parsed into a predicted aspect set $\hat{\mathcal{A}}_i^g$ and a generated description $\hat{d}_i^g$. Let $\mathcal{M}_{\text{reasoner}}$ denote the frozen Reasoner in QRPO. For each rollout, the Reasoner produces a predicted answer index $\hat{y}_i^g = \mathcal{M}_{\text{reasoner}}(q_i, O_i, \hat{d}_i^g, p_{\text{reasoner}})$, where p_{reasoner} is the reasoning prompt provided in Appendix B.3. Subsequently, QRPO assigns each rollout a composite reward from three components: format reward R_{fmt}, aspect selection reward R_{asp}, and response reward R_{rsp}.

3.3.1 REWARD DESIGN

Format reward. The Perceiver is required to produce outputs in the same structured format as the SFT target defined in Equation 4. The format reward is defined as

$$R_{\text{fmt},i}^g = \begin{cases} 1, & \text{if } o_i^g \text{ can be parsed into } \hat{\mathcal{A}}_i^g \text{ and } \hat{d}_i^g \text{ under the above format,} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

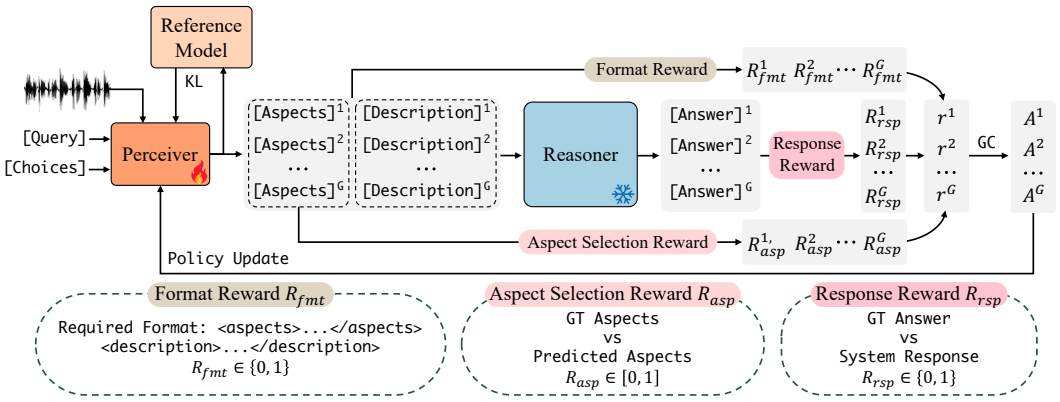

Figure 4: Overview of the QRPO procedure. GC and KL are the abbreviations of group computation and Kullback-Leibler divergence, respectively. The sample index subscript i is omitted for brevity.

Although the SFT stage already teaches the Perceiver to generate formatted outputs, we empirically find that incorporating a format reward during RL further improves training stability.

Aspect selection reward. To encourage query-oriented aspect selection, we measure the agreement between the predicted aspect set $\hat{\mathcal{A}}_i^g$ and the ground-truth aspect set $\mathcal{A}_i$ using the F1 score:

$$R_{\text{asp},i}^g = \text{F1}(\hat{\mathcal{A}}_i^g, \mathcal{A}_i) = \frac{2\left|\hat{\mathcal{A}}_i^g \cap \mathcal{A}_i\right|}{\left|\hat{\mathcal{A}}_i^g\right| + |\mathcal{A}_i|}. \tag{10}$$

Response reward. To align perception with downstream reasoning, the response reward is defined based on whether the Reasoner recovers the correct answer:

$$R_{\text{rsp},i}^g = \mathbb{I}[\hat{y}_i^g = y_i]\,. \tag{11}$$

This reward directly encourages the Perceiver to describe speech in a way that better supports downstream reasoning. The final reward for rollout g is a weighted combination of the three components:

$$r_i^g = \lambda_{\text{fmt}} R_{\text{fmt},i}^g + \lambda_{\text{asp}} R_{\text{asp},i}^g + \lambda_{\text{rsp}} R_{\text{rsp},i}^g, \tag{12}$$

where λ_{fmt}, λ_{asp}, and λ_{rsp} are hyperparameters that balance the three rewards.

3.3.2 Objective of QRPO

Following GRPO (Shao et al., 2024b), QRPO estimates the advantage of each rollout using group-relative normalization within the sampled outputs for the same input. For sample i, let μ_i and σ_i denote the mean and standard deviation of $\{r_i^g\}_{g=1}^G$. The advantage of rollout g is computed as

$$A_i^g = \frac{r_i^g - \mu_i}{\sigma_i + \delta}, \tag{13}$$

where δ is a small constant for numerical stability. QRPO then optimizes the Perceiver by maximizing the following objective:

$$\mathcal{J}_{\text{QRPO}}(\theta) = \mathbb{E}_{x_i \sim \mathcal{D}_{\text{PERE}}^{\text{strong}},\, \{o_i^g\}_{g=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|x_i)} \Bigg[\frac{1}{G} \sum_{g=1}^{G} \frac{1}{|o_i^g|} \sum_{\tau=1}^{|o_i^g|} \min\left(\rho_{i,\tau}^g A_i^g,\ \text{clip}(\rho_{i,\tau}^g, 1-\epsilon, 1+\epsilon)\, A_i^g\right) - \beta\, \mathbb{D}_{\text{KL}}[\pi_\theta \,\|\, \pi_{\text{ref}}] \Bigg], \tag{14}$$

where $x_i = (a_i, q_i, O_i)$ is the Perceiver input, $|o_i^g|$ is the token length of the serialized rollout g, $\rho_{i,\tau}^g = \pi_\theta(o_{i,\tau}^g \mid x_i, o_{i,<\tau}^g) / \pi_{\theta_{\text{old}}}(o_{i,\tau}^g \mid x_i, o_{i,<\tau}^g)$ is the importance sampling ratio at token position τ, ϵ is the clipping threshold, β is the KL penalty coefficient, and π_{ref} is the reference policy initialized from the SFT checkpoint.

4 EXPERIMENTAL SETUP

4.1 IMPLEMENTATION DETAILS

Models. The Perceiver is initialized from Qwen2.5-Omni-7B (Xu et al., 2025a). Qwen2-7B-Instruct (Yang et al., 2024a) is used as the Reasoner during RL training and is subsequently replaced by Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) for evaluation. Results of using different models as the Perceiver or the Reasoner are reported in Appendix C and Appendix D, respectively.

Optimization. For the SFT stage, the Perceiver is trained for 5 epochs with a per-device batch size of 16. For the RL stage, the Perceiver is trained for 1 epoch with a per-device batch size of 6 and gradient accumulation over 4 steps. In both stages, optimization is performed using AdamW (Loshchilov & Hutter, 2017a) with a learning rate of 1×10^{-6}, a cosine annealing schedule (Loshchilov & Hutter, 2017b), and a warmup ratio of 0.05. The Perceiver is trained in bfloat16 precision using DeepSpeed ZeRO-2 across 16 NVIDIA H100 GPUs. The Reasoner remains frozen throughout the process.

QRPO hyperparameters. During the RL stage, the Perceiver generates $G = 8$ rollouts for each input, with a maximum generation length of 512 tokens. Rollouts are sampled using a temperature of 1.2, top-p of 0.92, top-k of 20, and a repetition penalty of 1.05. The Reasoner is decoded using each checkpoint's official generation configuration. The reward weights are empirically set to $\lambda_{\text{fmt}} = 0.1$, $\lambda_{\text{asp}} = 0.2$, and $\lambda_{\text{rsp}} = 0.7$, while the KL regularization coefficient is fixed at $\beta = 0.05$.

4.2 TRAINING CORPORA

All training data are drawn from the SpeechCraft subset of AudioMCQ (He et al., 2026a; Jin et al., 2024). After the curation pipeline in Section 3.1, we obtain two partitioned corpora: $\mathcal{D}_{\text{PERE}}^{\text{weak}}$ for SFT training and $\mathcal{D}_{\text{PERE}}^{\text{strong}}$ for RL training, following the weak-to-strong training strategy in He et al. (2026a). Specifically, $\mathcal{D}_{\text{PERE}}^{\text{weak}}$ contains 74,475 samples, from which we randomly hold out 500 samples for validation. Since each sample is duplicated for the dual-objective SFT procedure in Section 3.2, the effective number of training and validation instances during the SFT training is doubled to 147,950 and 1,000, respectively. $\mathcal{D}_{\text{PERE}}^{\text{strong}}$ contains 124,051 samples, from which we randomly hold out 1,000 samples for validation during the RL training. During both the SFT and RL stages, we always provide candidate options to PERE. At evaluation time, whether candidate options are provided depends on their availability and on the evaluation setting.

4.3 EVALUATION BENCHMARKS

Conventional MCQ setting. We evaluate PERE on three commonly used multiple-choice question answering benchmarks: MMAU (Sakshi et al., 2025), MMSU (Wang et al., 2026b), and MMAR (Ma et al., 2026). Since our focus is on speech understanding, we evaluate only the speech subsets of MMAU and MMAR. In contrast, MMSU is specifically designed for spoken language understanding and is therefore evaluated in its entirety. Under this setting, candidate answer options are provided to the system, which is instructed to select the correct answer from the given choices. We report accuracy as the evaluation metric.

Open-ended setting. In practical applications, users typically provide only a speech signal and a query, without candidate answer options. To evaluate performance under a more realistic scenario, we conduct additional experiments on MMAU, MMSU, and MMAR, where candidate options are provided exclusively to the Reasoner and withheld from the Perceiver. This setting enables us to assess the Perceiver's behavior under realistic conditions, while still allowing the Reasoner to select from the available options, thereby facilitating straightforward accuracy computation. For a fair comparison with existing models, baseline SLMs are also restricted to speech and query inputs only; their outputs are subsequently passed to the same Reasoner used in PERE, together with the candidate options, to produce the final answer. Furthermore, we evaluate PERE on the `speech_QA` split of the chat subset in AIR-Bench (Yang et al., 2024b), a natively open-ended benchmark that does not provide candidate answer options. Instead of measuring accuracy, model outputs are evaluated by an LLM judge implemented with Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025), which assigns a score on a scale of 1 to 10. The judging prompt is provided in Appendix B.6.

Table 1: Speech understanding performance under the conventional MCQ and open-ended settings. Results for prior work under the conventional MCQ setting are taken from the original papers or the corresponding benchmark leaderboards. Results under the open-ended setting are obtained by evaluating the official checkpoints released by these models under our evaluation protocol.

Model	MMAU		MMSU			MMAR	AIR Bench
	Test-mini	Test	Per.	Rea.	Avg		
Conventional MCQ Setting							
Audio-CoT (Ma et al., 2025)	55.26	–	–	–	–	34.01	–
Audio-Reasoner (Xie et al., 2025)	66.07	62.53	–	–	–	32.99	–
Audio Flamingo 3 (Ghosh et al., 2026)	66.37	66.97	–	–	62.30	–	–
Baichuan-Omni (Li et al., 2025)	–	–	35.42	67.19	50.58	40.48	–
DIFFA (Zhou et al., 2026)	59.46	–	40.28	72.92	56.04	–	–
EvA (Xie et al., 2026)	68.47	–	47.52	75.41	62.24	63.01	–
GAMA (Ghosh et al., 2024)	12.91	16.97	–	–	–	27.89	–
Kimi-Audio (Ding et al., 2025)	62.16	56.57	43.52	76.03	59.28	–	–
Mellow (Deshmukh et al., 2026)	29.73	38.77	–	–	–	24.83	–
Qwen2-Audio-Instruct (Chu et al., 2024)	55.26	55.37	39.02	68.90	53.27	32.31	–
Qwen2.5-Omni-3B (Xu et al., 2025a)	–	–	42.37	72.76	56.83	53.74	–
Qwen2.5-Omni-7B (Xu et al., 2025a)	70.57	68.90	42.50	79.83	60.57	59.86	–
SALMONN (Tang et al., 2024)	26.43	28.77	–	–	–	34.69	–
PERE (Ours)	**74.47**	**69.73**	**48.01**	**80.36**	**63.16**	**64.93**	–
Open-Ended Setting							
Audio-Reasoner (Xie et al., 2025)	60.47	58.23	42.07	77.56	59.41	56.80	6.54
Audio Flamingo 3 (Ghosh et al., 2026)	52.85	50.37	44.57	78.02	60.86	55.44	7.35
Baichuan-Omni (Li et al., 2025)	54.65	48.93	43.02	71.82	57.10	47.96	7.27
DIFFA (Zhou et al., 2026)	44.14	40.00	40.48	58.16	49.04	31.97	5.94
GAMA (Ghosh et al., 2024)	27.03	25.57	31.04	47.27	39.58	22.11	4.22
Kimi-Audio (Ding et al., 2025)	56.46	51.87	44.52	74.47	59.08	46.26	6.81
Mellow (Deshmukh et al., 2026)	40.84	38.00	32.73	48.12	39.95	33.67	2.51
Qwen2-Audio-Instruct (Chu et al., 2024)	54.65	47.20	43.82	72.58	57.75	38.44	7.61
Qwen2.5-Omni-3B (Xu et al., 2025a)	55.56	51.23	45.51	73.95	59.28	44.90	7.27
Qwen2.5-Omni-7B (Xu et al., 2025a)	58.86	49.77	46.76	75.13	60.50	47.96	7.79
SALMONN (Tang et al., 2024)	30.75	29.87	28.31	52.42	40.05	39.46	3.00
PERE (Ours)	**65.47**	**60.97**	**48.57**	**80.60**	**64.14**	**57.14**	**8.32**

5 Experimental Results

In this section, we first compare PERE with existing SLMs and then conduct ablation studies on the key components of PERE. Additional experiments and analyses are provided in the Appendix.

5.1 Speech Understanding Performance

Conventional MCQ setting. As shown in the upper half of Table 1, PERE consistently outperforms existing SLMs across all benchmarks under the MCQ setting. Specifically, PERE achieves 74.47% on MMAU Test-mini and 69.73% on MMAU Test. On MMSU, PERE attains an average accuracy of 63.16%, with improvements observed on both the Perception and Reasoning subsets. On MMAR, PERE achieves 64.93%, surpassing all previous systems by a substantial margin.

Open-ended setting. The lower half of Table 1 reports the results under the open-ended setting, where the performance gap between PERE and existing methods becomes even more pronounced. Specifically, PERE achieves 65.47% on MMAU Test-mini and 60.97% on MMAU Test. On MMSU, PERE attains the highest average accuracy of 64.14%, while remaining highly competitive on MMAR with a score of 57.14%. On the open-ended AIR-Bench split, PERE achieves a score of 8.32, outperforming all existing systems. These results demonstrate that PERE generalizes effectively to realistic deployment scenarios where candidate answer options are unavailable.

Table 2: Ablation study of perception paradigms, SFT objectives, training stages, and QRPO reward components. Results on MMAU, MMSU, and MMAR are reported under the MCQ setting.

Variant	MMAU		MMSU			MMAR	AIR Bench
	Test-mini	Test	Per.	Rea.	Avg		
Implicit Perception							
Qwen2.5-Omni-7B	70.57	68.90	42.50	79.83	60.57	59.86	7.79
Qwen2.5-Omni-7B (Fine-tuned)	72.07	68.67	44.64	79.96	61.28	59.86	7.10
Query-Agnostic Perception							
Generic Captions + Reasoner	71.77	65.67	45.78	76.16	60.49	59.18	7.67
Query-Oriented and Reasoning-Aware Perception							
Qwen2.5-Omni-7B + Reasoner	65.77	65.87	45.59	74.36	59.59	50.34	8.06
+ Perception-only SFT	70.27	68.33	45.85	78.88	61.84	57.82	8.12
+ Answer-preserving SFT	71.17	69.33	47.01	78.36	62.16	59.52	8.13
+ QRPO w/o R_{fmt}	69.67	69.43	47.44	78.55	62.51	62.52	8.17
+ QRPO w/o R_{asp}	72.07	**70.20**	46.34	79.55	62.41	61.16	8.22
+ QRPO w/o R_{rsp}	72.07	69.43	46.13	78.67	61.90	60.82	8.16
+ QRPO (PERE)	**74.47**	69.73	**48.01**	**80.36**	**63.16**	**64.93**	**8.32**

5.2 Ablation Studies

Effectiveness of query-oriented and reasoning-aware perception. Table 2 compares three perception paradigms. Under implicit perception, to control for the influence of training data, we fine-tune Qwen2.5-Omni-7B on the same training data of PERE. However, it yields only modest improvements over the original model. Under query-agnostic perception, generic captions generated by Qwen3-Captioner (Xu et al., 2025b) are used as intermediate representations for the Reasoner. However, this approach does not consistently improve performance over the implicit baseline, suggesting that task-agnostic descriptions often fail to capture evidence critical for answering specific queries. In contrast, PERE employs a query-oriented and reasoning-aware perception paradigm, consistently achieving the strongest performance across all benchmarks.

Effectiveness of SFT and RL training stages. We progressively build up the full PERE pipeline to assess the contribution of each training stage. As shown in the last part of Table 2, starting from the untrained Perceiver paired with the Reasoner, direct coupling yields relatively weak performance. Introducing perception-only SFT improves the results; however, performance still lags behind the implicit perception baseline on both MMAU and MMAR. Adding answer-preserving SFT further enhances performance, suggesting that retaining the Perceiver's answer-generation capability during SFT helps mitigate degradation in speech understanding ability. Finally, applying QRPO after SFT yields the largest gains, leading to consistent and substantial improvements across all benchmarks.

Effectiveness of reward design. As shown in the last part of Table 2, removing the format reward R_{fmt} reduces training stability and leads to a noticeable performance drop on MMAU Test-mini. Removing the aspect selection reward R_{asp} or the response reward R_{rsp} degrades overall performance. The full QRPO objective with all three reward components achieves the best overall performance across all benchmarks, which demonstrates the effectiveness of our reward design.

6 Conclusion

We propose PERE, a modular Perceiver–Reasoner framework, together with QRPO, a reinforcement learning algorithm that enables query-oriented and reasoning-aware speech perception. Unlike implicit or query-agnostic perception, PERE explicitly determines what to perceive based on the query and expresses the perceived information in a form that facilitates downstream reasoning. Another key advantage of PERE is its plug-and-play Reasoner, which can be instantiated with arbitrary off-the-shelf LLMs, enabling direct access to state-of-the-art reasoning capabilities. Experiments on MMAU, MMSU, MMAR, and AIR-Bench demonstrate that PERE consistently outperforms existing speech language models under both conventional MCQ and more realistic open-ended settings.

REFERENCES

Siddhant Arora, Haidar Khan, Kai Sun, Xin Luna Dong, Sajal Choudhary, Seungwhan Moon, Xinyuan Zhang, Adithya Sagar, Surya Teja Appini, Kaushik Patnaik, et al. Stream RAG: Instant and accurate spoken dialogue systems with streaming tool usage. *arXiv preprint arXiv:2510.02044*, 2025.

Hongbo Bai, Yujin Zhou, Yile Wu, Chi-Min Chan, Pengcheng Wen, Kunhao Pan, Sirui Han, and Yike Guo. Glance-or-gaze: Incentivizing lmms to adaptively focus search via reinforcement learning. *arXiv preprint arXiv:2601.13942*, 2026.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Weidong Chen, Helen M. Meng, and Xixin Wu. Gogo: Group-wise granularity-ordered codec for stable and efficient speech generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Soham Deshmukh, Shuo Han, Hazim Bukhari, Benjamin Elizalde, Hannes Gamper, Rita Singh, and Bhiksha Raj. Audio entailment: Assessing deductive reasoning for audio understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23769–23777, 2025.

Soham Deshmukh, Satvik Dixit, Rita Singh, and Bhiksha Raj. Mellow: a small audio language model for reasoning. *Advances in Neural Information Processing Systems*, 38:49292–49332, 2026.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6288–6313, 2024.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations*, 2024.

Google DeepMind. Gemma 4. `https://ai.google.dev/gemma/docs/core/model_card_4`, 2026.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Xin Jin, Zhenguo Li, James T Kwok, and Yu Zhang. Reasoning-aligned perception decoupling for scalable multi-modal reasoning. *arXiv preprint arXiv:2506.04559*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi Liu, Zining Liang, Chunyat Wu, Qianhua He, Tan Lee, Xie Chen, Wei-Long Zheng, Weiqiang Wang, Mark D Plumbley, Jian liu, and Qiuqiang Kong. Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Xiang He, Chenxing Li, Jinting Wang, Yan Rong, Tianxin Xie, Wenfu Wang, Li Liu, and Dong Yu. Audio-deepthinker: Progressive reasoning-aware reinforcement learning for high-quality chain-of-thought emergence in audio language models. *arXiv preprint arXiv:2604.18187*, 2026b.

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

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

Tianle Liang, Yifu Chen, Shengpeng Ji, Yijun Chen, Zhiyang Jia, Jingyu Lu, Fan Zhuo, Xueyi Pu, Yangzhuo Li, and Zhou Zhao. Voxmind: An end-to-end agentic spoken dialogue system. *arXiv preprint arXiv:2604.15710*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017a.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017b.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng-Siong Chng, and Xie Chen. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–6, 2025.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *Advances in Neural Information Processing Systems*, 38, 2026.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.

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

Björn Schuller and Anton Batliner. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024a.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 16607–16629, 2024.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, et al. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*, 2025.

Jinchuan Tian, Sang gil Lee, Zhifeng Kong, Sreyan Ghosh, Arushi Goel, Chao-Han Huck Yang, Wenliang Dai, Zihan Liu, Hanrong Ye, Shinji Watanabe, Mohammad Shoeybi, Bryan Catanzaro, Rafael Valle, and Wei Ping. UALM: Unified audio language model for understanding, generation and reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Jinchuan Tian, Haoran Wang, Bo-Hao Su, Chien-yu Huang, Qingzheng Wang, Jiatong Shi, William Chen, Xun Gong, Siddhant Arora, Chin-Jou Li, et al. Bagpiper: Solving open-ended audio tasks via rich captions. *arXiv preprint arXiv:2602.05220*, 2026b.

Chaoren Wang, Heng Lu, Xueyao Zhang, Shujie Liu, Yan Lu, Jinyu Li, and Zhizheng Wu. Closing the modality reasoning gap for speech large language models. *arXiv preprint arXiv:2601.05543*, 2026a.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen M. Meng. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Dingdong Wang, Shujie Liu, Tianhua Zhang, Youjun Chen, Jinyu Li, and Helen M. Meng. Emotionthinker: Prosody-aware reinforcement learning for explainable speech emotion reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026c.

Yuanyuan Wang, Dongchao Yang, Yiwen Shao, Hangting Chen, Jiankun Zhao, Zhiyong Wu, Helen Meng, and Xixin Wu. Dualspeechlm: Towards unified speech understanding and generation via dual speech token modeling with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(40):33728–33736, 2026d.

Yucheng Wang, Jing Peng, Hanqi Li, Chenghao Wang, Wenming Tu, Yu Xi, Zhaokai Sun, Kai Yu, and Shuai Wang. Audio-mind: An auditable agentic framework for audio understanding. *arXiv preprint arXiv:2605.28480*, 2026e.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.

Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. Audio-thinker: Guiding large audio language model when and how to think via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33962–33970, 2026.

Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5187–5202, 2025.

Xinyuan Xie, Shunian Chen, Zhiheng Liu, Yuhao Zhang, Zhiqiang Lv, Liyin Liang, and Benyou Wang. Eva: An evidence-first audio understanding paradigm for lalms. *arXiv preprint arXiv:2603.27667*, 2026.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23829–23851, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

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*, 2025b.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.

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

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, 2024b.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB: Speech processing universal performance benchmark. In *Interspeech*, pp. 1194–1198, 2021.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. SALMONN-omni: A standalone speech LLM without codec injection for full-duplex conversation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Jiaming Zhou, Hongjie Chen, Shiwan Zhao, Jian Kang, Jie Li, Enzhi Wang, Yujie Guo, Haoqin Sun, Hui Wang, Aobo Kong, et al. Diffa: Large language diffusion models can listen and understand. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35112–35120, 2026.

Appendix

A Aspect Definition **14**

B Prompt Templates **15**

B.1 Aspect Selection Prompt 15

B.2 Description Extraction Prompt 15

B.3 Reasoner and Filter Prompt 16

B.4 Perceiver Prompt 17

B.5 Answer-preserving SFT Prompt 18

B.6 Judging Prompt 18

C Using Different Perceiver in PERE **18**

D Using Different Reasoner in PERE **19**

E Ablation Study on Data Strategy **20**

F Reward Curves in QRPO **21**

G Aspect Selection Accuracy and Description Quality **21**

H Experiment on Long Audio **23**

I Case Study **23**

I.1 Query-Oriented Aspect Selection 23

I.2 Comparison with Implicit and Query-Agnostic Perception 25

I.3 Error Analysis 27

J Limitations **28**

K The Use of Large Language Models **28**

A Aspect Definition

In PERE, we define a taxonomy of six perception aspects that cover the major perceptual dimensions of speech: $\mathcal{A} = \{\alpha^{\text{lex}}, \alpha^{\text{int}}, \alpha^{\text{emo}}, \alpha^{\text{pro}}, \alpha^{\text{flu}}, \alpha^{\text{spk}}\}$. The detailed definitions are as follows:

Speech Perception Aspects

1. Lexical & Semantic Content: key words, entities, literal meaning
2. Intent & Pragmatics: communicative goal, implied meaning grounded in speech
3. Emotion & Affect: emotion type and intensity
4. Prosody & Delivery: pitch, rhythm, stress, speaking rate
5. Fluency & Articulation: pauses, disfluencies, clarity
6. Speaker Traits: age, gender, accent, speaking style

B PROMPT TEMPLATES

B.1 ASPECT SELECTION PROMPT

The following prompt is used to implement the aspect selection function ϕ during data curation with Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025).

Aspect Selection Prompt

You are an expert in speech understanding and data annotation.

Your task is to decide which aspect IDs (from 1 to 6) are NECESSARY to answer the given multiple−choice speech question.

Available aspects:
1. Lexical & Semantic Content: key words, entities, literal meaning
2. Intent & Pragmatics: communicative goal, implied meaning grounded in speech
3. Emotion & Affect: emotion type and intensity
4. Prosody & Delivery: pitch, rhythm, stress, speaking rate
5. Fluency & Articulation: pauses, disfluencies, clarity
6. Speaker Traits: age, gender, accent, speaking style

Input fields:
[Question]
{question}
[/Question]

[Choices]
{choices}
[/Choices]

[Answer]
{answer}
[/Answer]

[CoT]
{cot}
[/CoT]

Instructions:
− Select the MINIMAL set of aspects required to answer this question correctly.
− Use only IDs from 1 to 6.
− Prefer fewer aspects when possible.
− Output IDs in ascending order.
− Do not include any explanation.

Output format (STRICT; no extra text):
<aspects>IDs of selected aspects separated by commas</aspects>

B.2 DESCRIPTION EXTRACTION PROMPT

The following prompt is used to implement the description extraction function ψ during data curation with Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025).

Description Extraction Prompt

You are an expert speech annotation assistant.

You will receive:
1) A multiple−choice speech question and options.
2) The correct answer.

3) A chain−of−thought (CoT), which may contain:
– descriptions of audio/speech signals,
– reasoning about the question text,
– references to final answer choice labels/options.

Task:
Extract and rewrite ONLY the information from the CoT that is directly grounded in the speech signal.
Do not add any information that is not stated in the CoT.

Keep:
– lexical content that is actually spoken in audio (keywords/entities/utterances),
– prosody, rhythm, intonation, stress, speaking rate,
– fluency/disfluency, pauses, articulation clarity,
– emotion/affect and intensity,
– speaker traits that are inferable from audio (e.g., accent, age/gender cues, style),
– pragmatic meaning only when inferable from spoken delivery/context in the audio.

Remove:
– question paraphrases and instructions,
– explicit option comparison logic tied to multiple−choice letters/numbers,
– statements like ”therefore answer is X”, ”option B is correct”,
– any text that depends on external world knowledge not present in audio evidence.

Output requirements:
– Write one concise paragraph in English.
– Focus on objective audio evidence.
– Do NOT mention ”question”, ”options”, ”answer”, ”correct choice”, or option IDs.
– Do NOT include XML tags inside the description itself.

Input:
[Question]
{question}
[/Question]

[Choices]
{choices}
[/Choices]

[Aspects]
{aspects}
[/Aspects]

[CoT]
{cot}
[/CoT]

[Answer]
{answer}
[/Answer]

Output format (STRICT; no extra text):
<description>your cleaned speech−signal−only description</description>

B.3 Reasoner and Filter Prompt

The following prompt is used by the reasoning-sufficiency filter $\mathcal{M}_{\text{filter}}$ and the Reasoner $\mathcal{M}_{\text{reasoner}}$ to generate the final answer based on the speech description. Note that the blue-highlighted portion is omitted during open-ended evaluation.

Reasoner and Filter Prompt

You are an expert in reasoning over speech descriptions.

You are given:
1) A description of the speech
2) A user query

Your task is to directly answer the user query based on the information in the description.

Speech description:
{description}

User query:
{query}

Options:
– {o_1}
– {o_2}
– {o_3}
– {o_4}

Choose exactly one of the listed options. Reply with only the chosen option text; match the wording of that option as closely as possible.

B.4 Perceiver Prompt

The following prompt is used by the Perceiver during perception-only SFT, QRPO training, and evaluation. Note that the blue-highlighted portion is omitted during open-ended evaluation.

Perceiver Prompt

Your task is to determine which aspects of speech the query depends on, and then extract detailed evidence from those aspects to support answering the query.

Given a user query, do:
1. Identify what information is REQUIRED to answer the query.
2. Determine the MINIMAL set of aspects that directly contain this information.
3. Provide concise, concrete evidence based ONLY on those aspects.

Available aspects:
1. Lexical & Semantic Content: key words, entities, literal meaning
2. Intent & Pragmatics: communicative goal, implied meaning grounded in speech
3. Emotion & Affect: emotion type and intensity
4. Prosody & Delivery: pitch, rhythm, stress, speaking rate
5. Fluency & Articulation: pauses, disfluencies, clarity
6. Speaker Traits: age, gender, accent, speaking style

IMPORTANT:
– You are NOT a question–answering system
– You do NOT respond to the user's query
– Your role is to provide a precise, evidence–based description of the speech
– Include ONLY information that is necessary for answering the query

Output Format:
<aspects>selected aspects</aspects><description>relevant evidence only</description>

User query:
{query}

Please choose the answer from the following options: [{o_1}, {o_2}, {o_3}, {o_4}].

B.5 ANSWER-PRESERVING SFT PROMPT

The following prompt is used by the Perceiver during answer-preserving SFT.

Answer-preserving SFT Prompt

Answer the user query based on the speech.
Wrap your answer strictly between the tags <answer> and </answer>.
Example: <answer>Your answer here</answer>

User query:
{query}

Please choose the answer from the following options: [{o_1}, {o_2}, {o_3}, {o_4}].

B.6 JUDGING PROMPT

The following prompt is used by an LLM judge based on Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) to evaluate system outputs on the AIR-Bench dataset. The `meta_info` field is provided by AIR-Bench. To mitigate position bias, we evaluate each sample twice: once with the ground-truth answer assigned to Assistant 1 and the system output assigned to Assistant 2, and once with the assignments reversed. The final score is computed by averaging the two corresponding ratings.

Judging Prompt

You are a helpful and precise assistant for checking the quality of the answer.

[Detailed Audio Description]
{meta_info}

[Question]
{query}

[The Start of Assistant 1s Answer]
{answer} or {gt_answer}
[The End of Assistant 1s Answer]

[The Start of Assistant 2s Answer]
{gt_answer} or {answer}
[The End of Assistant 2s Answer]

[System]
We would like to request your feedback on the performance of two AI assistants in response to the user question and audio description displayed above. AI assistants are provided with detailed audio descriptions and questions. Please rate the helpfulness, relevance, accuracy, and comprehensiveness of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance. Please output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space.

C USING DIFFERENT PERCEIVER IN PERE

To assess whether the gains of PERE depend on a specific backbone, we keep the Reasoner fixed to Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) and instantiate the Perceiver with three representative SLMs: Qwen2.5-Omni-7B (Xu et al., 2025a), Qwen2.5-Omni-3B (Xu et al., 2025a), and Qwen2-Audio-Instruct (Chu et al., 2024). All Perceivers are trained with the same PERE pipeline and evaluated under the same protocol as Table 1. As shown in Tables 3, 4, and 5, applying PERE consistently improves performance over the corresponding direct-response SLM across MMAU, MMSU, MMAR, and AIR-Bench under both the conventional MCQ and open-ended settings. No-

tably, fine-tuning each SLM on the same training data used by PERE yields only modest gains and still falls short of the full PERE pipeline, suggesting that the performance improvements are not merely attributable to additional training data, but rather to the proposed query-oriented and reasoning-aware perception paradigm. Taken together, these results demonstrate that the benefits of PERE are not tied to a single SLM architecture, highlighting the robustness and generalizability of the PERE framework.

Table 3: Performance of PERE with Qwen2.5-Omni-7B as the Perceiver.

Variant	MMAU		MMSU			MMAR	AIR Bench
	Test-mini	Test	Per.	Rea.	Avg		
Conventional MCQ Setting							
Qwen2.5-Omni-7B	70.57	68.90	42.50	79.83	60.57	59.86	–
Qwen2.5-Omni-7B (Fine-tuned)	72.07	68.67	44.64	79.96	61.28	59.86	–
Qwen2.5-Omni-7B + Reasoner	65.77	65.87	45.59	74.36	59.59	50.34	–
+ Perception-only SFT	70.27	68.33	45.85	78.88	61.84	57.82	–
+ Answer-preserving SFT	71.17	69.33	47.01	78.36	62.16	59.52	–
+ QRPO (PERE)	**74.47**	**69.73**	**48.01**	**80.36**	**63.16**	**64.93**	–
Open-Ended Setting							
Qwen2.5-Omni-7B	58.86	49.77	46.76	75.13	60.50	47.96	7.79
Qwen2.5-Omni-7B (Fine-tuned)	60.66	52.00	46.24	73.76	59.56	48.64	7.10
Qwen2.5-Omni-7B + Reasoner	59.16	57.53	46.23	75.54	60.98	48.30	8.06
+ Perception-only SFT	62.46	58.73	47.51	78.01	62.81	52.38	8.22
+ Answer-preserving SFT	62.46	59.00	46.97	79.18	63.12	55.44	8.23
+ QRPO (PERE)	**65.47**	**60.97**	**48.57**	**80.60**	**64.14**	**57.14**	**8.32**

Table 4: Performance of PERE with Qwen2.5-Omni-3B as the Perceiver.

Variant	MMAU		MMSU			MMAR	AIR Bench
	Test-mini	Test	Per.	Rea.	Avg		
Conventional MCQ Setting							
Qwen2.5-Omni-3B	–	–	42.37	72.76	56.83	53.74	–
Qwen2.5-Omni-3B (Fine-tuned)	63.56	64.40	44.40	73.22	57.87	52.59	–
Qwen2.5-Omni-3B + Reasoner	61.56	59.57	41.01	66.47	53.39	46.94	–
+ Perception-only SFT	66.46	65.53	41.15	75.23	57.66	51.70	–
+ Answer-preserving SFT	66.97	65.67	42.21	76.76	58.93	53.40	–
+ QRPO (PERE)	**69.97**	**66.77**	**44.97**	**77.02**	**60.50**	**55.10**	–
Open-Ended Setting							
Qwen2.5-Omni-3B	55.56	51.23	45.51	73.95	59.28	44.90	7.27
Qwen2.5-Omni-3B (Fine-tuned)	57.36	54.17	46.29	78.83	62.98	50.20	7.88
Qwen2.5-Omni-3B + Reasoner	54.95	51.13	42.51	73.53	57.55	45.92	7.73
+ Perception-only SFT	57.36	55.77	44.85	75.98	60.07	45.67	7.88
+ Answer-preserving SFT	58.86	56.87	47.33	78.61	61.50	48.64	8.00
+ QRPO (PERE)	**59.46**	**57.37**	**48.72**	**79.26**	**63.51**	**53.40**	**8.02**

D USING DIFFERENT REASONER IN PERE

A key advantage of PERE is its modular design: the Reasoner can be replaced with arbitrary off-the-shelf LLMs without retraining the Perceiver, thereby leveraging the strongest available reasoning capabilities as LLMs continue to advance. To validate this flexibility, we fix the Perceiver to the default PERE checkpoint trained with Qwen2.5-Omni-7B (Xu et al., 2025a) and swap only the Reasoner among several representative LLMs, including Gemma-4-26B-A4B-it (Google DeepMind,

Table 5: Performance of PERE with Qwen2-Audio-Instruct as the Perceiver.

Variant	MMAU		MMSU			MMAR	AIR Bench
	Test-mini	Test	Per.	Rea.	Avg		
Conventional MCQ Setting							
Qwen2-Audio-Instruct	55.26	55.37	39.02	68.90	53.27	32.31	–
Qwen2-Audio-Instruct (Fine-tuned)	62.16	63.17	46.24	71.83	57.58	44.62	–
Qwen2-Audio-Instruct + Reasoner	57.36	45.87	30.46	57.46	43.53	35.71	–
+ Perception-only SFT	62.46	61.87	46.23	73.42	59.81	45.09	–
+ Answer-preserving SFT	65.47	64.27	46.08	75.45	60.30	47.28	–
+ QRPO (PERE)	**68.17**	**66.73**	**48.06**	**76.74**	**61.96**	**50.34**	–
Open-Ended Setting							
Qwen2-Audio-Instruct	54.65	47.20	43.82	72.58	57.75	38.44	7.61
Qwen2-Audio-Instruct (Fine-tuned)	54.35	53.37	45.18	73.57	58.09	46.95	7.83
Qwen2-Audio-Instruct + Reasoner	48.65	45.60	38.09	66.80	51.97	35.37	6.59
+ Perception-only SFT	56.16	54.63	40.92	72.42	55.41	45.20	7.49
+ Answer-preserving SFT	59.76	56.63	42.91	76.34	58.74	46.26	7.80
+ QRPO (PERE)	**60.36**	**58.27**	**46.12**	**78.13**	**61.62**	**50.68**	**7.97**

Table 6: Performance of PERE with different Reasoners.

Reasoner	MMAU		MMSU			MMAR	AIR Bench
	Test-mini	Test	Per.	Rea.	Avg		
Conventional MCQ Setting							
Gemma-4-26B-A4B-it	**75.37**	69.63	47.52	80.08	63.01	61.90	–
Llama-3.2-3B-Instruct	64.56	63.03	45.48	76.40	60.45	54.42	–
Qwen2-7B-Instruct	72.67	66.70	44.46	78.07	60.72	58.84	–
Qwen3-30B-A3B-Instruct-2507	74.47	**69.73**	**48.01**	**80.36**	**63.16**	**64.93**	–
Open-Ended Setting							
Gemma-4-26B-A4B-it	61.86	55.20	**52.45**	**81.43**	**67.61**	54.76	8.04
Llama-3.2-3B-Instruct	57.66	57.27	45.81	73.69	59.30	50.34	7.48
Qwen2-7B-Instruct	62.46	58.17	44.31	77.25	60.26	51.02	7.89
Qwen3-30B-A3B-Instruct-2507	**65.47**	**60.97**	48.57	80.60	64.14	**57.14**	**8.32**

2026), Llama-3.2-3B-Instruct (Grattafiori et al., 2024), Qwen2-7B-Instruct (Yang et al., 2024a), and Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025). All results are reported under the same evaluation protocol as Table 1. As shown in Table 6, overall performance varies with the choice of Reasoner, confirming that downstream reasoning quality directly affects the final system output. Notably, replacing the default Reasoner with Gemma-4-26B-A4B-it yields stronger results on MMAU under the MCQ setting and MMSU under the open-ended setting, further improving upon the already competitive performance achieved with Qwen3-30B-A3B-Instruct-2507. The above results indicate that, once optimized, the Perceiver generalizes effectively to previously unseen LLMs at inference time without requiring additional retraining. Furthermore, the plug-and-play nature of PERE highlights its long-term scalability: as more capable LLMs become available, the framework can directly leverage their enhanced reasoning abilities without requiring any modification or retraining of the perception module.

E Ablation Study on Data Strategy

We ablate two key design choices in our data curation pipeline: the reasoning-sufficiency filter and the weak-to-strong training data partition. Specifically, the reasoning-sufficiency filter $\mathcal{M}_{\text{filter}}$ retains only samples whose generated descriptions are sufficient for a separate model to recover the

Table 7: Ablation study on training data strategies. Results on MMAU, MMSU, and MMAR are reported under the MCQ setting.

Variant	MMAU		MMSU			MMAR	AIR Bench
	Test-mini	Test	Per.	Rea.	Avg		
Reasoning-Sufficiency Filter							
w/o $\mathcal{M}_{\text{filter}}$	71.47	69.53	46.45	79.75	62.58	62.14	8.20
w/ $\mathcal{M}_{\text{filter}}$ (PERE)	**74.47**	**69.73**	**48.01**	**80.36**	**63.16**	**64.93**	**8.32**
Training Data Partition							
Mixed weak and strong	73.27	69.10	47.62	79.71	63.15	62.59	8.26
Weak-to-strong (PERE)	**74.47**	**69.73**	**48.01**	**80.36**	**63.16**	**64.93**	**8.32**

correct answer. This step ensures that the Perceiver is supervised with descriptions that genuinely support downstream reasoning, rather than descriptions that omit critical speech evidence or contain spurious cues. As shown in Table 7, removing this filter and training on all curated samples degrades performance across benchmarks, indicating that not every automatically extracted description provides equally reliable supervision. By filtering out low-quality samples, PERE improves the quality of the training signal and better aligns the Perceiver's learning objective with downstream reasoning.

Regarding the training data partition, we compare our default weak-to-strong strategy with a mixed alternative that uses both weak- and strong-audio-contribution samples for both SFT and RL. As shown in the bottom of Table 7, under the mixed setting, overall performance drops slightly across all benchmarks, with a more noticeable gap on MMAU and MMAR. This observation is consistent with AudioMCQ (He et al., 2026a), which first proposed partitioning training data by audio-contribution difficulty and demonstrated the benefits of a weak-to-strong training strategy over uniformly mixing all samples. We attribute this to the differing roles of the two training stages: the weak subset provides relatively accessible supervision for learning the basic perception behavior and output format during SFT, while the strong subset exposes the Perceiver to samples that more heavily depend on speech evidence during RL, encouraging query-oriented and reasoning-aware perception.

F Reward Curves in QRPO

Figure 5 illustrates the reward trajectories during QRPO training for PERE with different Perceivers: Qwen2-Audio-Instruct, Qwen2.5-Omni-3B, and Qwen2.5-Omni-7B. For each model, we report the total reward along with the three reward components used in QRPO: the format reward R_{fmt}, aspect selection reward R_{asp}, and response reward R_{rsp}.

As shown in Figure 5, the format reward remains consistently high throughout training (≥ 0.9), indicating that the structured output format learned during SFT is effectively preserved during RL training. In contrast, both the aspect selection reward and the response reward exhibit a steady upward trend over the course of training, suggesting that the Perceiver progressively improves its ability to identify query-oriented perceptual aspects and generate descriptions that better support downstream reasoning. As these reward components improve, the total reward increases correspondingly, demonstrating stable and effective optimization of QRPO across different Perceiver backbones.

G Aspect Selection Accuracy and Description Quality

The Perceiver in PERE is trained to perform query-oriented and reasoning-aware perception, which requires determining both what information should be perceived to answer a given query and how the perceived evidence should be described to facilitate downstream reasoning. To evaluate these capabilities more directly, we design a dedicated experiment to assess aspect selection accuracy and description quality. Specifically, we use the validation part of $\mathcal{D}_{\text{PERE}}^{\text{strong}}$ (described in Section 4.2) for evaluation. For aspect selection, we compute the macro-F1 score between the aspects predicted by the model and the ground-truth aspects, which reflects the model's ability to identify the perceptual aspects most relevant to the given query. For description quality, we feed the model-generated



Figure 5: Reward curves during QRPO training.

descriptions into the same Reasoner used in PERE to obtain the final answers. The accuracy between these predicted answers and the ground-truth answers serves as a proxy for description quality, measuring how effectively the generated descriptions support downstream reasoning.

The visualizations of aspect selection accuracy and description quality are presented in Figure 6. Specifically, we compare three model variants: the base model without PERE-style training, the model after PERE's SFT stage, and the fully trained PERE model after both SFT and QRPO training. This comparison allows us to examine how each training stage contributes to query-oriented aspect selection and the generation of reasoning-aware descriptions. As shown in Figure 6, regardless of which model is used as the Perceiver, both aspect selection accuracy and description quality improve substantially after the SFT stage. These metrics are further enhanced following QRPO training. These results suggest that the original base models possess limited capability for query-oriented and reasoning-aware perception. In contrast, after SFT and QRPO training, the PERE Perceiver becomes significantly more effective at identifying query-relevant perceptual aspects and generating high-quality descriptions, thereby providing stronger support for downstream reasoning.

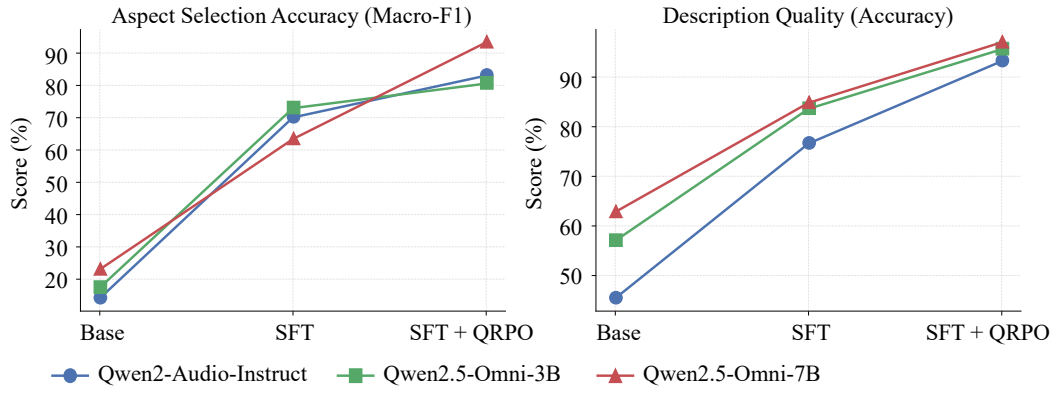

Figure 6: Visualization of aspect selection accuracy and description quality.

H EXPERIMENT ON LONG AUDIO

Existing speech understanding benchmarks mostly contain relatively short clips. To evaluate whether query-oriented and reasoning-aware perception remains effective when the input speech is longer, we construct **LongSpeechBench-20s**, a subset of 424 speech samples whose duration is at least 20 seconds. Specifically, we collect all speech samples with duration ≥ 20 s from the MMAU Test-mini, MMSU, and MMAR benchmarks used in our main evaluation, yielding 135, 147, and 142 samples, respectively. All models are evaluated under the open-ended setting.

As shown in Table 8, speech understanding becomes substantially more challenging when processing longer audio inputs. Nevertheless, PERE achieves the highest average accuracy of 60.14%, outperforming all existing SLMs. These results confirm that PERE can effectively avoid distraction from irrelevant content and selectively focus on the portions of long speech that are most relevant to the given query. Although the absolute performance on LongSpeechBench-20s is lower than that on the full benchmarks reported in Table 1, PERE consistently maintains its advantage over existing SLMs, demonstrating the effectiveness of our proposed PERE for long-form speech understanding.

Table 8: Results on LongSpeechBench-20s under the open-ended setting. The benchmark contains 424 speech samples (≥ 20 s) drawn from MMAU Test-mini, MMSU, and MMAR.

Method	MMAU	MMSU	MMAR	Avg
Audio-Reasoner (Xie et al., 2025)	58.52	62.59	53.52	58.25
Audio-Flamingo-3 (Ghosh et al., 2026)	56.30	57.82	51.41	55.19
Baichuan-Omni-1.5 (Li et al., 2025)	54.07	56.46	48.59	53.07
DIFFA (Zhou et al., 2026)	48.15	43.54	34.51	41.98
GAMA (Ghosh et al., 2024)	23.70	29.25	19.72	24.29
Kimi-Audio (Ding et al., 2025)	57.04	52.38	47.18	52.12
Mellow (Deshmukh et al., 2026)	33.33	36.73	32.39	34.20
Qwen2-Audio-7B-Instruct (Chu et al., 2024)	57.78	54.42	39.44	50.47
Qwen2.5-Omni-3B (Xu et al., 2025a)	54.81	48.30	43.66	48.82
Qwen2.5-Omni-7B (Xu et al., 2025a)	58.52	54.42	45.77	52.83
SALMONN (Tang et al., 2024)	56.30	41.50	38.73	45.28
PERE (Ours)	**59.26**	**66.67**	**54.23**	**60.14**

I CASE STUDY

The preceding experiments quantify the effectiveness of PERE through performance metrics. To provide further insight into the model's behavior, we present case studies under both the conventional MCQ setting, where candidate answer options are available, and the open-ended setting, where only the user query is provided. Through these case studies, we demonstrate that PERE dynamically selects different perceptual aspects according to the query (see Section I.1), generates descriptions that better facilitate downstream reasoning (see Section I.2), and, owing to its modular design, makes perception errors interpretable and traceable to downstream reasoning failures (see Section I.3).

I.1 QUERY-ORIENTED ASPECT SELECTION

The following examples use the same speech clip paired with two different queries, under the MCQ and open-ended settings respectively. They illustrate how the Perceiver selects different perception aspects and generates distinct descriptions depending on the information required by each query.

Case 1a: Query-Oriented Aspect Selection (MCQ setting)

Speech: Multi-speaker conversational dialogue

Query A: Identify the emotion of each speaker.

Options (Query A):

A. third speaker show neutral and joy
B. first speaker shows sadness, neutral, and joy; others are neutral
C. first and second speakers show joy, others are neutral
D. third and fourth speakers show joy, others are neutral

Ground Truth (Query A): A. third speaker show neutral and joy

Selected Aspects (Query A): 3 (Emotion & Affect), 4 (Prosody & Delivery)

Description (Query A): The first speaker, a young female, delivers her lines with a normal pitch and high volume, conveying sadness through a subdued tone and deliberate pacing, followed by neutral delivery in her subsequent utterances. The second speaker, an older female, speaks with a calm, steady pitch and moderate volume, maintaining a neutral tone throughout. The third speaker, a young male, exhibits a consistently high pitch and energetic delivery, particularly in his repeated exclamation of "Great, look!" with clear joy and excitement. The fourth speaker, another young male, speaks with a neutral pitch and moderate volume, delivering his lines without emotional inflection.

Predicted Answer (Query A): A. third speaker show neutral and joy

Query B: What workplace is mentioned in the dialogue?

Options (Query B):

A. Central Perk
B. Szechuan Dragon
C. Moondance Diner
D. Irving Benson Memorial Highway

Ground Truth (Query B): B. Szechuan Dragon

Selected Aspects (Query B): 1 (Lexical & Semantic Content)

Description (Query B): A young female speaker delivers the utterance "Because I don't work at the Szechuan Dragon," with clear articulation and a rising intonation on "Szechuan Dragon," followed by a brief pause and a shift in pitch on "OK," indicating a conversational tone. The name is spoken with emphasis on the syllable "Szechuan" and a slightly elongated vowel sound, suggesting it is a proper noun.

Predicted Answer (Query B): B. Szechuan Dragon

Case 1b: Query-Oriented Aspect Selection (Open-ended setting)

Speech: Single-speaker utterance ("I see the problem now")

Query A: What is the speaker's gender and speaking rate?

Ground Truth (Query A): A young female speaker with a slow speaking rate

Selected Aspects (Query A): 6 (Speaker Traits)

Description (Query A): A young female speaker with a high-pitched voice delivers the utterance "I see the problem now" at a slow speaking rate, with clear articulation and natural emotional expression, maintaining a steady rhythm and moderate intonation throughout.

Predicted Answer (Query A): A young female speaker with a slow speaking rate

Query B: What is emphasized by the stress in this sentence?

Ground Truth (Query B): The specific issue being discussed

Selected Aspects (Query B): 1 (Lexical & Semantic Content), 4 (Prosody & Delivery)

Description (Query B): A young female speaker delivers the utterance "I see the problem now" with a slow, deliberate pace, clear articulation, and a rising intonation on "the problem," followed by a brief pause before "now," which is spoken with a steady, confident tone.

Predicted Answer (Query B): The specific issue being discussed

Analysis. In Case 1a, Query A asks about the speakers' emotions, prompting the Perceiver to select aspects 3 and 4 (Emotion & Affect and Prosody & Delivery) and produce a speaker-wise affect profile for all four participants. In contrast, Query B, posed on the same dialogue, focuses on lexical content. Accordingly, the Perceiver selects only aspect 1 (Lexical & Semantic Content) and identifies "Szechuan Dragon" from the utterance. Case 1b exhibits the same behavior for a single-speaker utterance. For Query A, which asks about the speaker's gender and speaking rate, the Perceiver selects aspect 6 (Speaker Traits) and describes the speaker as "a young female" with "a slow speaking rate." For Query B, which concerns prosodic emphasis, the Perceiver instead activates aspects 1 and 4, highlighting the emphasis on "the problem" rather than the speaker's identity. These examples demonstrate that the Perceiver dynamically selects perceptual aspects according to the information required by the query, rather than relying on a fixed perception strategy, regardless of whether candidate answer options are provided.

I.2 Comparison with Implicit and Query-Agnostic Perception

The following examples compare PERE with an implicit direct-response SLM and a query-agnostic captioning baseline on the same samples, under the MCQ and open-ended settings respectively.

Case 2a: Comparison with Baselines (MCQ setting)

Speech: Single-speaker utterance ("I see the problem now")

Query: What is emphasized by the stress in this sentence?

Options:

A. The specific issue being discussed
B. The timing of the problem
C. The realization or understanding
D. The solution has been found

Ground Truth: A. The specific issue being discussed

Implicit Perception (Qwen2.5-Omni-7B):

Predicted Answer: C. The realization or understanding

Query-Agnostic Perception (Generic Caption + Reasoner):

Generic Caption: The audio clip begins with a brief, low-frequency thump, likely caused by the microphone being handled or bumped, immediately followed by a soft, low-pitched inhalation as the speaker prepares to speak. The recording environment is acoustically dry and intimate, with no background noise, reverberation, or environmental sounds, indicating a small, well-dampened room and close microphone placement. The speaker, a young adult female with a General American English accent, delivers the phrase "I see the problem now" in a slow, deliberate, and contemplative manner. Her voice is soft, breathy, and low-pitched, with a gentle, downward inflection on "now" that conveys a sense of resignation and quiet realization. The phrase is articulated clearly, with the word "the" pronounced with a voiced dental fricative, and the overall delivery is marked by a subtle, melancholic tone. The recording is of high fidelity, with no distortion, clipping, or electronic artifacts, and the only audible sounds are the speaker's voice and the initial thump. The clip ends abruptly after the final word, with no fade-out or lingering sound.
In summary, the audio presents a solitary, emotionally nuanced moment: a young woman, in a quiet and private setting, softly acknowledges a realization with a tone of resignation and introspection. The technical and acoustic qualities reinforce the sense of intimacy and emotional weight, suggesting the recording is intended for personal reflection or as a candid, unembellished expression of thought.

Predicted Answer: C. The realization or understanding

Query-Oriented and Reasoning-Aware Perception (PERE):

Selected Aspects: 1 (Lexical & Semantic Content), 4 (Prosody & Delivery)

Description: A young female speaker with a high-pitched, normal-volume voice delivers the utterance "I see the problem now" at a natural pace, emphasizing the phrase "the problem" with clear stress and a slight pause before it.

Predicted Answer: A. The specific issue being discussed

Case 2b: Comparison with Baselines (Open-ended setting)

Speech: Single-speaker utterance ("The first successful organ transplant was . . . ")

Query: What organ was transplanted in the procedure mentioned by the speaker?

Ground Truth: Kidney

Implicit Perception (Qwen2.5-Omni-7B):

Predicted Answer: The speaker did not specify which organ was transplanted.

Query-Agnostic Perception (Generic Caption + Reasoner):

Generic Caption: The audio clip begins with a single female speaker, whose voice is clear, steady, and delivered in a neutral, professional tone. She speaks in standard American English, with no regional accent, and her articulation is precise and formal, suggesting a prepared statement. The speaker states, "The first successful organ transplant was a breathtaking achievement in medical history." Her delivery is measured and evenly paced, with a slight emphasis on the word "breathtaking," which is pronounced with a subtle upward inflection, conveying a sense of awe and significance. The phrase "in medical history" is spoken with a slightly lower pitch, marking the conclusion of the statement. The recording is of high fidelity, with no background noise, music, or environmental sounds, and the voice is centered in the stereo field, indicating a close-mic setup in a controlled, acoustically treated environment. The audio ends abruptly, with no fade-out or trailing silence, suggesting it is an excerpt from a longer recording.
In summary, the clip features a professionally recorded, high-quality excerpt of a female speaker delivering a concise, formal statement about the historical significance of the first successful organ transplant. The delivery is neutral and precise, with a subtle emotional emphasis on the word "breathtaking," and the absence of any background or environmental sounds underscores the controlled, studio-like setting. The content and style suggest the audio is intended for educational, documentary, or broadcast purposes, and the abrupt ending indicates it is a segment from a larger production.

Predicted Answer: Liver

Query-Oriented and Reasoning-Aware Perception (PERE):

Selected Aspects: 1 (Lexical & Semantic Content), 2 (Intent & Pragmatics)

Description: The speaker states, "The first successful organ transplant was a breathtaking achievement in medical history." The utterance focuses on a landmark "organ transplant" procedure but does not name the organ explicitly. The speaker highlights the historical significance of this pioneering transplant as a major milestone in medicine, framing the procedure itself as the focal referent of the statement.

Predicted Answer: Kidney

Analysis. In Case 2a, both the implicit SLM and the query-agnostic baseline predict option C ("The realization or understanding"), likely due to the phrase "I see . . . now," which suggests a moment of realization. Moreover, the generic caption emphasizes the recording environment and overall delivery but fails to identify the prosodic stress on "the problem," which is crucial for answering the query. In contrast, PERE selects aspects 1 and 4 and explicitly highlights "the problem" as the stressed constituent in its description, enabling the Reasoner to correctly select option A. In Case 2b, the answer is not explicitly stated in the speech and instead requires reasoning based on background knowledge. The implicit SLM refuses to answer, stating that the speaker did not specify which organ was transplanted. The generic caption summarizes the utterance but also includes a substantial amount of query-irrelevant information, which misleads the Reasoner into making an incorrect inference. In contrast, PERE generates a focused description centered on the key phrase "organ transplant," providing the critical evidence needed for the Reasoner to infer the correct answer, "Kidney." These examples demonstrate that query-oriented and reasoning-aware descriptions provide more task-relevant evidence than implicit responses or query-agnostic captions, particularly

when the answer depends on combining speech content with external knowledge rather than relying solely on information explicitly stated in the speech.

I.3 Error Analysis

The following examples present failure cases in which an incomplete or incorrect perception output prevents the Reasoner from arriving at the correct answer.

Case 3a: Perception Error Propagation (MCQ setting)

Speech: Multi-speaker conversational dialogue (four speakers)

Query: How many speakers are speaking in this conversation?

Options:

A. one
B. two
C. four
D. three

Ground Truth: C. four

Selected Aspects: 1 (Lexical & Semantic Content)

Description: A male speaker delivers the utterance "What? You're coming on to the entire room!" followed by a brief pause and a shift to a female speaker saying "I'm Chandler" and "I'm Marjorie", indicating two distinct speakers with different vocal characteristics and turns in the conversation.

Predicted Answer: B. two

Failure Cause: The Perceiver describes only two speakers and omits the remaining participants; the Reasoner cannot recover the correct count from the incomplete description.

Case 3b: Perception Error Propagation (Open-ended setting)

Speech: Multi-speaker conversational dialogue (four speakers)

Query: How many speakers are speaking in this conversation?

Ground Truth: four

Selected Aspects: 1 (Lexical & Semantic Content)

Description: A young female speaker delivers the lines "What? You're coming on to the entire room" and "I'm Chandler", followed by a male speaker saying "I'm Marjorie" and "I talk in my sleep", indicating two distinct voices with clear vocal differences.

Predicted Answer: two

Failure Cause: The Perceiver collapses a multi-speaker exchange into a two-speaker interaction; once this under-count is written into the description, the Reasoner has no basis for inferring additional participants.

Analysis. In Case 3a, the Perceiver's description includes only a male speaker and a female speaker while omitting the other two participants in the four-speaker dialogue. Given this incomplete description, the Reasoner reasonably predicts option B ("two"), whereas the ground-truth answer is option C ("four"). In Case 3b on the same dialogue under the open-ended setting, the Perceiver similarly collapses the interaction into two voices and the Reasoner consequently reports "two" instead of the correct count of "four". In both cases, the error stems from incomplete perception rather than incorrect downstream reasoning: the Perceiver produces an inaccurate description, and the Reasoner has no textual basis to infer additional speakers or to verify or correct the Perceiver's output, as it operates solely on the generated text. This modular design makes the failure readily interpretable by directly comparing the Perceiver's description with the underlying speech content.

J LIMITATIONS

PERE is currently designed for speech understanding and may not generalize directly to other audio modalities such as music or environmental sounds, where the perception aspect taxonomy would need to be redefined. As a modular framework, PERE requires sequential inference by both the Perceiver and the Reasoner, resulting in higher latency and memory consumption compared with end-to-end SLMs. In addition, our training corpus is derived from a single source dataset and remains relatively limited in scale. Finally, since the Reasoner operates solely on the description produced by the Perceiver, errors or omissions in perception may propagate to the final response and cannot be corrected by downstream reasoning.

K THE USE OF LARGE LANGUAGE MODELS

Large language models were employed exclusively as auxiliary tools to edit and polish text written by the authors. Their usage was limited to improving clarity, grammar, and style of expression. No part of the research ideation, methodology, analysis, or results relied on LLMs.