Speech-Hands: A Self-Reflection Voice Agentic Approach to Speech Recognition and Audio Reasoning with Omni Perception

Zhen Wan[1,2] **Chao-Han Huck Yang**[1] **Jinchuan Tian**[1,3] **Hanrong Ye**[1] **Ankita Pasad**[1]
Szu-wei Fu[1] **Arushi Goel**[1] **Ryo Hachiuma**[1] **Shizhe Diao**[1] **Kunal Dhawan**[1]
Sreyan Ghosh[1] **Yusuke Hirota**[1] **Zhehuai Chen**[1] **Rafael Valle**[1]
Chenhui Chu[2] **Shinji Watanabe**[3] **Boris Ginsburg**[1] **Yu-Chiang Frank Wang**[1]
[1]NVIDIA [2]Kyoto University [3]Carnegie Mellon University
Corresponding authors: zhenwan@nlp.ist.i.kyoto-u.ac.jp; hucky@nvidia.com

Abstract

We introduce a voice-agentic framework that learns one critical omni-understanding skill: knowing when to trust itself versus when to consult external audio perception. Our work is motivated by a crucial yet counterintuitive finding: naively fine-tuning an omni-model on both speech recognition and external sound understanding tasks often degrades performance, as the model can be easily misled by noisy hypotheses. To address this, our framework, Speech-Hands, recasts the problem as an explicit self-reflection decision. This learnable reflection primitive proves effective in preventing the model from being derailed by flawed external candidates. We show that this agentic action mechanism generalizes naturally from speech recognition to complex, multiple-choice audio reasoning. Across the OpenASR leaderboard, Speech-Hands consistently outperforms strong baselines by 12.1% WER on seven benchmarks. The model also achieves 77.37% accuracy and high F1 on audio QA decisions, showing robust generalization and reliability across diverse audio question answering datasets. By unifying perception and decision-making, our work offers a practical path toward more reliable and resilient audio intelligence. [1]

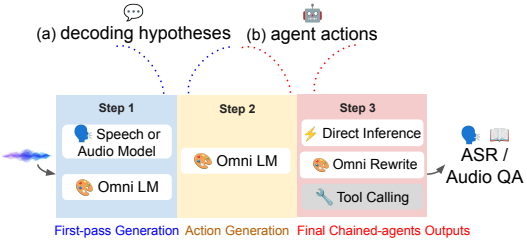


Figure 1: *Speech-Hands* acts as a dynamic orchestrator that predicts a special action token to govern its cognitive strategy for ASR and multi-domain audio reasoning.

1 Introduction

Omni-modal models (Xie and Wu, 2024; OpenAI et al., 2024; Xu et al., 2025a; Li et al., 2025b) that jointly process audio and text have unified a range of audio understanding tasks, including automatic speech recognition (ASR), temporal sound event reasoning, and knowledge-heavy question answering. However, human perception is not naturally perfect at understanding acoustic patterns across different resolutions at soundscape (Calcus, 2024). For example, while professional speech interpreters could produce superior ASR transcriptions, this specialized ability does not guarantee a comparable aptitude for understanding animal sounds or complex music (Galantucci et al., 2006).

Inspired by developmental psychology (Selman and Byrne, 1974), we draw a parallel to the human capacity for self-reflection, where children mature from a purely egocentric viewpoint to a stage of *self-reflective perspective-taking*, which serves the critical ability to "step outside" one's own thoughts, evaluate one's beliefs against others', and importantly, to recognize the boundaries of one's own knowledge. In contrast, current models often operate egocentrically, implicitly trusting their internal perception without the capacity to critically assess its reliability or seek external assistance when necessary. We aim to instill a form of computational self-reflection (Nelson, 1990) into an omni-modal agent, designing a collaborative framework that explicitly reasons about when to trust its own perception, when to defer to an expert, and even when to utilize tools.

We frame these voice understanding models not just as passive predictors, but as agentic that have access to multiple internal and external information sources, and must decide how to best use them. For such an agent, a central decision-making challenge arises (Lebiere and Anderson, 2011): should it rely on its own auditory perception, or consult external suggestions, such as ASR alternatives or other per-

[1]Project page, interactive demo, and code: https://YukinoWan.github.io/Speech-Hands/

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 43124–43142

ceptual consultants? Prior work, such as ASR and large language model (LLM) cascaded approach of Generative Error Correction (GER) (Yang et al., 2023a; Lin et al., 2025), as shown in Figure 5, sidesteps this agentic dilemma entirely. By operating only on text hypotheses without access to the original audio, these methods are fundamentally non-agentic; they cannot weigh internal perception against external advice because they have no internal perception to begin with. Our preliminary experiments reveal that naively combining modalities often degrades performance, as the model struggles to resolve conflicts between its own perception and flawed external suggestions (Kaiser et al., 2021). Without a mechanism to decide which source to trust, the model is easily confused.

To address this, we introduce *Speech-Hands*, a learnable framework that instantiates self-reflection as a core control primitive. As illustrated in Figure 1, our agent operates as a dynamic orchestrator. It begins by aggregating multi-source decoding hypotheses (a) from the first-pass generation. Instead of blindly fusing inputs, the agent critically evaluates them to predict an explicit agent action (b) during the action generation phase. This control token effectively dicates the model's cognitive strategy: trigerring fast direct inference when confidence is high (selecting whether its internal perception or external perceptions), engaging in omni rewrite over available evidence, or initiating tool calling when special utilities are required. In this work, we will focus on the actions of direct inference and omni rewrite, leaving the tool calling action in future work. This approach unifies transcription and reasoning under a single, controllable framework that knows when to trust, when to rethink, and when to ask for help.

1.1 Preliminary: The Surprising Failure of Multimodal Correction with Omni-LM

A natural hypothesis is that providing an omni model with both audio and text hypotheses during supervised fine-tuning (SFT) should enhance GER performance. We tested this assumption by fine-tuning Qwen2.5-Omni (Xu et al., 2025b) to correct N-best hypotheses ($N = 5$) from Whisper-v2-large (Radford et al., 2023) on OpenASR datasets.

Figure 2 results, however, are surprisingly negative. As shown in our preliminary study (Table 3), this naive SFT approach consistently degrades performance across seven ASR benchmarks, yielding a higher Word Error Rate (WER) than either of the baseline models alone.

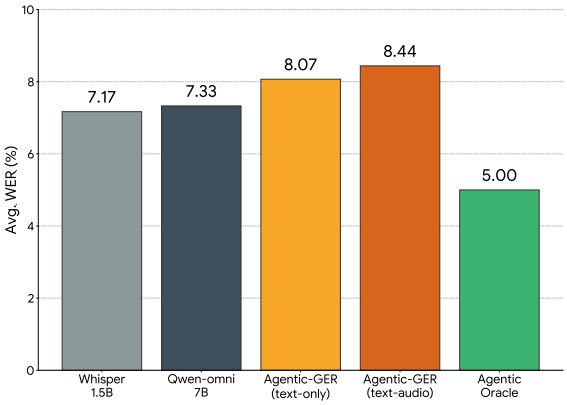


Figure 2: Preliminary results on the cascaded agentic Qwen-omni baseline for generative error correction (GER) with supervised fine-tuning show that both text-only and text-audio GER degrade ASR performance, where the best ASR and LLM combination achieves a low agentic output oracle of 5% WER.

To verify that this degradation was not due to suboptimal prompting, we extensively tested varying instructions during SFT, which aims to emphasize internal audio perception, external transcripts, or a balanced fusion. However, as shown in Table 1, all prompting setups failed to recover performance (e.g., WER increased to 8.52%–9.05% compared to baselines).

AVG. WER	(a)	(b)
Emphasize internal	-	8.63
Emphasize external	8.58	8.67
Emphasize audio	9.02	9.05
Balanced	8.44	8.52

Table 1: Prompt ablation results in preliminary SFT. (a) audio + external whisper 5-best and (b) audio + internal 1-best + external whisper 5-best. Details in Appendix A.3

Furthermore, our zero-shot analysis reveals that the base model lacks intrinsic arbitration capabilities: its decisions are highly sensitive to prompt wording rather than ground truth, often collapsing into trivial heuristics (as shown in Table 2 and details in Appendix A.4).

These findings demonstrate a fundamental flaw in the naive omni-LM fusion approach: without a mechanism to adjudicate between its own perception and potentially flawed external advice, the omni-model is easily confused, often amplifying hallucinations or overcorrections as shown in the ASR failure case (Appendix A and Section 6.3).

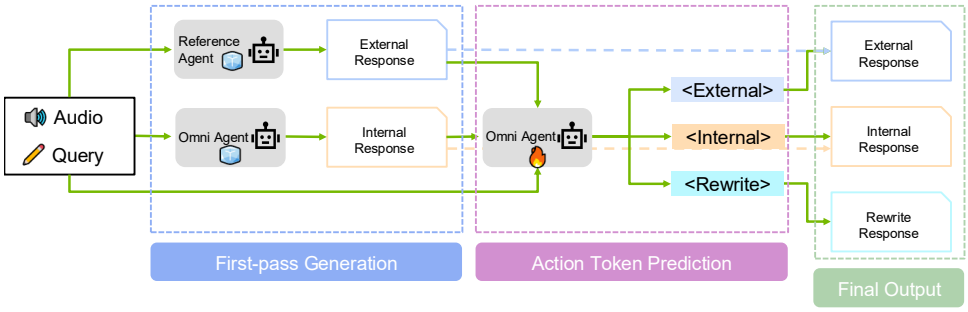


Figure 3: Overview of our proposed Self-Reflection Multimodal GER framework. A special token is generated at the beginning to decide whether to use audio perception (i.e., transcription hypotheses or caption) or not.

Table 2: Confusion matrix of zero-shot decisions under different prompting strategies. The results show that the model's arbitration is highly sensitive to prompt wording rather than the ground truth correctness (Oracle), often collapsing into trivial heuristics.

Ground Truth	Internal-biased Prompt		External-biased Prompt		Balanced Prompt	
	Pred. Int	Pred. Ext	Pred. Int	Pred. Ext	Pred. Int	Pred. Ext
Oracle Internal	0.83	0.17	0.35	0.65	0.68	0.32
Oracle External	0.71	0.29	0.14	0.86	0.34	0.66

This provides strong motivation for a more principled mechanism that allows the model to learn when and how to incorporate external information.

2 Related Work

2.1 Voice Retrieval and Agentic Framework

Voice retrieval has become a useful component in augmenting audio tasks such as captioning (Koizumi et al., 2020; Zhao et al., 2023), audio-to-text generation (Huang et al., 2023), dialogue system (Chen et al., 2025), and music generation (Gonzales and Rudzicz, 2024). Recently, voice retrieval modules have been incorporated into multimodal agents (Yang et al., 2023b; Zhang et al., 2024; Wang et al., 2025; Wan et al., 2025), providing access to external memory or specialized tools across modalities.

Yet, even in these agentic frameworks, retrieval is often treated as an auxiliary enhancer, rather than as a distinct source of information. When the retrieved content diverges from the internal prediction of the model, current systems lack principled mechanisms for arbitration.

2.2 Omni-Modality and Self-Reflection in Multimodal Models

By weaving together text, vision, and audio into a single fabric of understanding, these systems begin to approach the fluidity of human perception (Xu et al., 2025b,c; Goel et al., 2024; Abouelenin et al., 2025). Modality-specific reflection methods (Hu et al., 2025) suggest that introspection within each sensory channel can partially bridge these gaps, aligning representations with quiet precision.

Recent works also introduce explicit self-reflection (Renze and Guven, 2024; Madaan et al., 2023) into multimodal reasoning (Cheng et al., 2024; Fang et al., 2025). The self-reflected video reasoner (Song et al.) iteratively critiques its own visual understanding to reinforce policy stability, while other efforts call for reflective checks against overconfidence and modality neglect (Yang et al., 2025b). Still, these mechanisms operate *after* perception, which treats reflection as a corrective mirror once fusion has already occurred.

Rather than reflecting on the output, our Speech-Hands framework reflects on the act of perception itself. It learns an action mechanism that decides, in real time, whether to trust its own *"ears"* or the *"words"* of others. We aim to discover self-reflection from a post-hoc repair strategy into a preemptive act of perceptual discernment toward an early glimmer of meta-cognition within multimodal understanding.

3 Methodology

We present **Speech-Hands**, a learnable omni-agentic framework for audio understanding and reasoning (Figure 3). It enables a multimodal language model to explicitly choose a special token: trusting its own internal perception or defer to external hypotheses, enables efficient **Fast Inference**, while rewriting a new response engages the **Omni Rewrite** process for deeper reasoning. This token is generated during inference and guides the downstream generation process, allowing interpretable and agentic decision-making.

3.1 Task Formulation

We formulate a unified self-reflection agent that generalizes across both speech recognition and audio question answering. Given an input audio A, an optional query Q, the agent first generates its own response H_{omni}, and then combined with an external response H_{ext} provided by an external model.

Next, rather than fusing these sources implicitly as normal GER researches, SpeechHands introduces a learned policy to explicitly choose among them. Our agent model first emits a special action token from the set $\{\texttt{<internal>}, \texttt{<external>}, \texttt{<rewrite>}\}$ to indicate whether to trust itself or rely on external hypothesis or even whether rewriting a new response after rethinking the audio task and all input resources for the final answer (GER). This self-reflection decision is made based on the full context $(A, Q, H_{\text{omni}}, H_{\text{ext}})$, and the selected action conditions the final generation.

To supervise the self-reflection mechanism, we construct ground-truth action token labels on the whole training dataset by comparing the performance of internal, external, and GER predictions. We leverage detailed strategies for action token construction in ASR and audio QA.

3.2 Action Token Construction for ASR

In the ASR setting, we use WER as a pointer to decide our actions. For each audio example, we first prompt the omni model to generate the transcript T_{int} and in parallel leverage an ASR model to predict the external T_{ext}, we then let omni model to generate again based on both the audio and the external T_{ext} to acquire the GER prediction T_{ger}. Next, we compute the WER between the ground-truth transcript T_{gt} and each of the three candidates. If T_{int} is the same as T_{gt} ($WER = 0$) or has the lowest WER, the label is assigned as <internal>, this is to encourage the model to trust itself when it can successfully solve the problem, otherwise <external> if T_{ext} performs best, or <rewrite> if T_{ger} performs best.

3.3 Action Token Construction for Audio QA

Unlike the ASR setting where token selection is based on fine-grained WER scores, Audio QA presents a discrete supervision signal: each prediction is either correct or incorrect. For each instance consisting of an audio segment A, a question Q, and answer choices $\mathcal{C}$, we first prompt the omni model to produce an internal prediction c_{int} based on (A, Q). In parallel, we obtain an external prediction c_{ext} from an audio reasoning model.

We then compare both predictions against the ground-truth answer c^*. If $c_{\text{int}} = c^*$, we assign the label <internal>, encouraging self-reliance when the model performs well. If c_{int} fails but $c_{\text{ext}} = c^*$, we assign <external> to delegate control. Otherwise, when both predictions are incorrect, the label is set to <rewrite>, signaling a need to re-evaluate the question with all available context.

However, this binary decision process introduces instability during training. External predictions can be stochastic. Therefore, repeated sampling may yield different answers, especially under high uncertainty or directly change another external model can also lead to a different accuracy. This stochasticity makes the decision boundary between <external> and <rewrite> inherently less robust.

To mitigate this, we adopt a multiple decoding-based strategy: for each example, we sample the external model five times and collect their predicted choices. If the majority of predictions match the ground-truth answer, we assign <external>; otherwise, we assign <rewrite>. This approach stabilizes supervision by reducing the variance in external outputs and yields more reliable action labels.

3.4 Prompt Formatting and Training

Subsequently, each training instance is formatted as a single target string consisting of the decision token followed by the final target transcript or answer, e.g. <rewrite> + ground-truth transcription. This unified string allows the model to learn not only how to generate the task output but also how to decide the action before generation. We adopt an instruction-style prompt to guide the model to make decisions, below is the prompt template for ASR:

```
You are an omni-agent for speech
understanding with access to three
inputs:
(1) The original audio;
(2) Five transcription hypotheses from
another ASR system (external);
(3) Your own first-pass transcription
(internal).


Your task is to:
- First decide whether your internal
transcription is reliable.
- If yes, output <internal> and your
transcription.
- If the external system is more reliable,
output <external> and use one of its
```

```
hypotheses.
- Otherwise, output <rewrite> and
generate a new answer using both sources
and the audio.
```

For Audio QA, the prompt tempalte is shown in Appendix B. During training, we optimize a single cross-entropy loss over the concatenated target sequence, which jointly supervises the action token and the subsequent prediction. Concretely, the model first predicts the action token and then continues decoding the target transcript or answer; both parts contribute to the same loss. This end-to-end objective encourages the model to internalize the mapping from multimodal evidence to action choice and to generate the corresponding output under that decision.

3.5 Agentic Inference via Action Tokens

At inference time, the model receives the same multimodal inputs as during training. The model performs a two-stage decoding process: it first emits an action token: `<internal>`, `<external>`, or `<rewrite>`. The decoding process is subject to decide which information source to prioritize, and then generates the final output accordingly. This explicit agentic self-reflection mechanism offers interpretability and control over how the model balances internal perception and external knowledge. It enables direct analysis (*e.g.*, F1 score) of when the model relies on its own understanding, consultants from external systems, or synthesizes a new response. The unified prediction format ensures that the model not only learns what to generate, but also which to trust across both speech recognition and audio reasoning tasks.

4 Experimental Setup

4.1 Datasets

Speech Recognition. We use seven representative datasets covering a range of domains, styles, and noise conditions from OpenASR leaderboard: AMI (Carletta, 2007) (meeting speech), Tedlium (Hernandez et al., 2018) (TED talks), GigaSpeech (Chen et al., 2021) (large-scale podcasts and YouTube-style speech), SPGISpeech (O'Neill et al., 2021) (long-form read speech), VoxPopuli (Wang et al., 2021) (English subset of multilingual political recordings), and LibriSpeech (Panayotov et al., 2015) (clean and noisy audiobook speech). For baseline fine-tuning and prompt-based GER, we use all available training sets. For our proposed method, unless otherwise specified (`w/ FULL Datasets`), we train on at most 20,000 examples per dataset, this is due to the limitation of heavy computation requirements when doing inference on the whole training set for internal, external and GER (token distribution is discussed in 6.1).

Audio Reasoning. We evaluate on the multi-domain audio question-answering benchmark (Yang et al., 2025a) (MD-Audio), which consists of multiple-choice questions grounded in real audio clips. This benchmark includes three complementary subsets that probe different reasoning capabilities. While previous audio reasoning benchmarks such as MMAU (Sakshi et al., 2024a) and MMAR (Ma et al., 2025) provide only MMLU-style test sets, MD-Audio releases both training and development sets, enabling evaluation of the proposed trainable agentic framework, as shown in Table 9. Detailed dataset descriptions can be found in Appendix D. For each sample, we construct inputs as described in Section 3, including the original audio, a first-pass internal prediction from Omni model, one external hypothesis and GER result (for ASR).

5 Results

5.1 Training Details

All experiments are conducted using the Qwen2.5-Omni model. We extend its tokenizer to include three special action tokens, `<internal>`, `<external>` and `<rewrite>`, used during both training and inference. Models are trained using the standard supervised fine-tuning (SFT) objective with a cross-entropy loss. We train for 5 epochs with fp16. The batch size is set to 64, and the learning rate is initialized at 1e-4 with cosine decay. All experiments adopt greedy decoding.

5.2 Baselines

External ASR models: We include three high-performing supervised speech recognition models as external references: Whisper-v2-large (Radford et al., 2022), Canary-1B-v2, and Parakeet-TDT-0.6B-v3 (Sekoyan et al., 2025). These systems operate in a closed transcription setting and provide non-generative references.

External audio QA model: For multi-choice audio QA, we include Audio Flamingo 3 (Goel et al., 2025), a speech-language model with audio frontend capabilities serving as a strong baseline.

We also include the latest model in comparison,

Dataset	AMI	Tedlium	Gigspeech	Spgispeech	VoxPopuli	Libri-clean	Libri-other	avg. WER ↓
ASR model or Omni-LLM								
Whisper-v2-large	16.88	4.32	11.45	3.94	7.57	2.91	5.15	7.17
Canary-1b-v2	19.80	4.78	11.66	3.08	6.35	1.73	3.17	7.22
Parakeet-tdt-0.6b-v3	12.69	4.90	12.24	3.16	6.48	1.89	3.37	6.68
Qwen2.5_omni	19.77	5.17	11.26	4.58	6.59	2.09	3.85	7.33
Phi-4-MM	11.69	**2.90**	**9.78**	3.13	**5.93**	1.68	3.83	6.14
Gemini-2-Flash	21.58	3.01	10.71	3.82	7.89	2.49	5.84	8.56
GPT-4o-voice	57.76	5.79	13.64	5.66	10.83	3.48	7.97	15.76
Qwen2.5-Omni: Cascaded								
GER: ⇒ Whisper-v2-large	23.44	6.15	12.15	3.94	7.53	2.97	4.89	8.44
GER: ⇒ canary	24.58	6.38	12.43	4.02	7.72	3.05	5.01	8.74
GER: ⇒ parakeet	22.91	6.09	12.10	3.98	7.49	2.92	4.84	8.33
Qwen2.5-Omni: Parallel								
Speech-Hands ⇌ whisper-v2	15.03	4.45	12.37	3.01	6.49	1.86	3.46	6.67
Speech-Hands ⇌ canary	15.29	4.21	10.87	**2.17**	5.96	**1.61**	**3.07**	6.17
Speech-Hands ⇌ parakeet	**11.20**	4.37	11.10	2.26	6.02	1.67	3.18	**5.69**

Table 3: WER (%) results across 7 datasets, with the average WER shown in the rightmost column. Speech-Hands training significantly outperforms both baseline systems (ASR model, Qwen) and prior cascaded GER setups.

Model / Setting	Bio-acoustic	Soundscape	Complex QA	avg. Acc. ↑
Audio LM or Omni Model				
Gemini-2-Flash	42.03	46.34	59.89	56.61
Qwen2.5-Omni	47.32	56.32	59.89	57.87
AudioFlamingo 3 (AF3)	71.88	57.31	81.26	74.49
Qwen2.5-Omni Baselines				
+ SFT with official training data	78.13	34.65	76.61	63.13
+ GRPO with official training data	78.09	39.43	79.12	65.54
+ GRPO with external audio data (Li et al., 2025a)	62.32	**72.10**	82.15	75.10
GER: ⇒ AF3 (cascaded agentic)	76.29	52.02	77.48	68.93
Qwen2.5-Omni: Speech-Hands				
⇌ (parallel agentic): SFT with official training data	67.86	58.29	83.34	75.75
⇌ (parallel agentic) + majority sampling	**81.25**	59.4	**85.7**	**77.37**

Table 4: AudioQA and acoustic content reasoning accuracy (%) across knowledge-intensive bioacoustic QA (Sayigh et al., 2016), multi-sound-object soundscapes, and MMAU-style (Sakshi et al., 2024a) complex audio QA tasks.

e.g., Phi-4-MM (Microsoft et al., 2025), Gemini-2-Flash, GPt-4o-voice in ASR evaluations.

5.3 ASR Results

Table 3 presents WER results across seven diverse datasets, and we find that: (1) Speech-Hands outperforms all baselines. Furthermore, our approach with strong ASR models (canary and parakeet) achieves the lowest WER, even with only 20k training examples, outperforming both ASR models or Omni-LLMs; (2) The promtp GER over Whisper lags significantly behind token-based methods. This underscores the importance of explicit control via action tokens rather than relying solely on natural-language prompts; (3) While pre-trained models like Whisper and Qwen perform well on curated datasets such as LibriSpeech-clean, their performance degrades significantly on conversational benchmarks like AMI. Notably, despite Qwen's relatively weaker base ASR performance, our framework enhances its generalization to the extent that it surpasses stronger baselines such as Phi-4-MM on the average performance, demonstrating the stability and transferability of Speech-Hands on both clean and noisy datasets.

5.4 AudioQA Results

Table 4 reports accuracy on each sub-task and overall average accuracy: (1) Our final setup (Speech-Hands + majority sampling) achieves the highest average accuracy (77.37%), outperforming all baselines and pre-trained models. It performs particularly well on Complex QA (85.70%) and Bioacoustics QA (81.25%), indicating its ability to handle both abstract reasoning and fine-grained audio patterns; (2) Standard supervised fine-tuning (SFT) and prompt-based GER exhibit mixed results. SFT achieves good accuracy on Bioacoustics (78.13%) but fails on Soundscapes (34.65%). Prompt-based GER also fails on both Soundscape and Complex

Dataset	AMI	Tedlium	Gigaspeech	SPGIspeech	Voxpopuli	Libri-clean	Libri-other
Training Distribution							
<internal>	67.95%	86.48%	87.76%	96.25%	93.73%	98.96%	98.96%
<external>	31.01%	11.18%	11.8%	3.64%	6.09%	0.96%	0.96%
<rewrite>	1.04%	2.34%	0.44%	1.21%	0.18%	0.1%	0.1%
Test Distribution							
<internal>	70.28%	83.57%	85.94%	95.42%	92.68%	98.92%	98.75%
<external>	26.91%	15.12%	12.41%	3.81%	6.47%	0.98%	1.08%
<rewrite>	2.27%	1.31%	1.65%	0.77%	0.85%	0.1%	0.17%
<internal> on Test							
Precision	0.85	0.63	0.77	0.89	0.85	0.88	0.83
Recall	0.78	0.72	0.90	0.94	0.90	0.99	0.99
F1	0.81	0.67	0.83	0.91	0.87	0.94	0.9
<external> on Test							
Precision	0.88	0.96	0.83	0.82	0.71	0.81	0.72
Recall	0.89	0.81	0.78	0.76	0.60	0.73	0.75
F1	0.89	0.88	0.80	0.79	0.65	0.77	0.74
<rewrite> on Test							
Precision	0.62	0.33	0.32	0.52	0.50	0.0	0.0
Recall	0.24	0.21	0.05	0.36	0.21	0.0	0.0
F1	0.39	0.28	0.08	0.43	0.33	0.0	0.0

Table 5: Training distribution and test-time F1 scores for Speech-Hands' action tokens.

QA compared with flamingo 3 baseline. These results highlight the robustness of our agentic framework in diverse audio reasoning settings.

6 Additional Analysis

6.1 Accuracy of Action Token Prediction

We analyze the model's ability to correctly emit the three action tokens of <internal>, <external>, and <rewrite>, in the ASR setting under the Action Tokens + Whisper configuration. These tokens interpret whether the model is making correct decisions to guide its final prediction.

Table 5 shows both the training distribution and the test-time precision, recall, and F1 scores. The distribution highlights a strong internal bias: across all datasets, <internal> dominates, exceeding 95% in Libri-clean, Libri-other, spgispeech, and Voxpopuli. In contrast, <external> is sparsely supervised, where below 1% in Librispeech and <rewrite> is extremely rare everywhere, often less than 2%. This imbalance poses a natural challenge for learning reliable action.

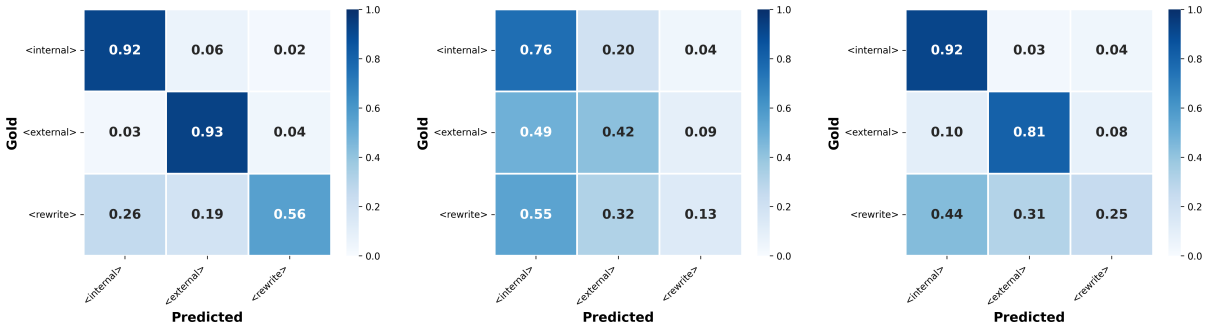
Despite this skew, the model demonstrates robust performance for the two tokens. On test data, <internal> predictions achieve F1 scores above 0.8 on most datasets (0.91 on spgispeech, 0.94 on Libri-clean, 0.90 on Libri-other), indicating that the model can reliably recognize when its own decoding is sufficient. Even for the much rarer <external> token, the model attains high F1 scores, showing strong generalization of deferring to external hypotheses despite limited supervision.

The <rewrite> token proves to be the most challenging, with F1 scores below 0.4 in all but one dataset and zero in Librispeech, where positive training examples are extremely rare. A closer examination reveals that precision consistently exceeds recall, indicating that when the model does emit <rewrite>, its decision is generally correct but under-triggered in the omni model. This suggests a cautious yet reasonably reliable rewrite detector, whose coverage could be further improved through targeted data augmentation.

Overall, these results validate the effectiveness of the proposed agentic action: even under heavy class imbalance, the model learns to accurately identify when to trust its own predictions versus when to consult external information. The main bottleneck remains the <rewrite> case, suggesting that richer sampling or augmentation strategies may be needed to stabilize this decision in future work.

6.2 Confusion Analysis In AudioQA

We analyze the confusion matrix over the three subsets under the w/ multiple sampling setup (the oracle token distribution is shown in Table 6). As shown in Figure 4, the confusion between <internal> and <external> remains relatively low, suggesting that the model can effectively dis-



(a) Bio-acoustic QA (b) Soundscape QA (c) Complex QA

Figure 4: Confusion matrices of Speech-Hands' agentic action execution for audio QA and reasoning three subsets on (a) bio-acoustic QA, (b) temporal and sound event QA, and (c) complex audio information QA.

Subset	<in>/<ex>/<re>
Bio-acoustic QA	106/75/43
Soundscape QA	343/185/81
Complex QA	978/555/100

Table 6: Oracle Statistics of the three DCASE2025 AudioQA test sets used in our experiments.

tinguish between them, especially in the Complex QA subset. However, the Soundscape subset shows slightly increased overlap, possibly due to the difficulty of Soundscape compared with other parts, also their original performances (Qwen 56.32 v.s. Flamingo 57.31) are quite closed, leading to much smaller sets of `<external>`. Besides, the highest confusion still lies with the `<rewrite>` class. As summarized in Table 9, the number of training tokens labeled as `<rewrite>` is significantly smaller across all subsets. Such sparsity likely limits the model's tendency to generalize the rewriting behavior during generation, but when it generates `<rewrite>`, the accuracy is quite high and robust as shown in the Figure. These findings highlight that even with imbalanced actions in training, the current F1 scores can still reliably improve the final performance, emphsizing the effectiveness of action tokens.

6.3 Category Analysis and Case Study

To provide systematic insight beyond individual cases, during our prelimiary, we categorize three primary failure modes observed in ASR when Speech-Hands is absent. These categories further motivate our design choice of action tokens: (1) **External-induced Misguidance:** the internal omni-model is correct, but the model is misled by an error-heavy external hypothesis. This is the most damaging failure mode and directly motivates the action-token arbitration design; (2) **Overcorrection:** LLMs tend to "complete" disfluent or partially heard speech, introducing hallucinated insertions or semantic expansions; (3) **Undercorrection in Dual-Failure:** when both internal and external predictions are wrong, the omni model often selects one erroneous hypothesis instead of generating a rewrite, due to the lack of an explicit rewrite mechanism. These observations align with the high-precision behavior of the `<rewrite>` token and explain the improvements: the model selectively activates `<rewrite>` only when necessary.

We present illustrative examples (more in Appendix E and F) to show how the model makes selective use of internal, external, and rewrite hypotheses across modalities. In the AUDIOQA CASE, both the internal and external models predict "B. Thunderstorm", likely influenced by surface acoustic features such as low-frequency rumbling. The rewrite path, however, generates "C. Forest fire", which aligns with the ground truth, demonstrating the strong influence of `<rewrite>` that even two models have the same first-pass prediction.

In contrast, the ASR CASE reveals a different decision dynamic. Although the rewrite produces a more fluent variant, the model opts to retain the baseline hypothesis, judging the original phrase as sufficiently accurate. This indicates the model's ability to avoid overcorrection issue in prior GER researches when input ambiguity is low. These examples underscore the benefits of explicit action tokens: the model can either rely on internal or external model predictions or revise them when necessary, yielding both flexibility and robustness in audio tasks.

AudioQA Inference-Time Case Study

Q. Based on the audio, which natural phenomenon could be occurring?
A. Earthquake B. Thunderstorm
C. Forest fire D. Snowstorm

Internal pred: B. Thunderstorm
External pred: B. Thunderstorm

Pred: <rewrite> C. Forest fire (✓)

ASR Inference-Time Case Study

Internal: you in the way marguerite but how
External: you ll in the way marguerite but how
Rewrite: you are in the way marguerite but how

Pred: <internal> you in the way marguerite but how (✓)

7 Conclusion

In this work, we proposed a learnable voice-agentic framework Speech-Hands for teaching omni models when to trust itself versus when to consult external audio perception. By casting the problem with explicit <internal>, <external>, and <rewrite> action tokens, our experimental results across AudioQA and ASR benchmarks demonstrate strong performance improvements beyond strong baselines, especially when direct finetuning and GER training fail, Speech-Hands can still robustly generate the best prediction.

This framework also benifits the interpretability in analysis, the model achieves high F1 scores for both <internal> and <external> tokens, even under imbalanced training conditions. While the <rewrite> token is rarer, its precision notably exceeds recall, indicating that the model can accurately identify necessary rewrites when it does trigger them. Overall, our method offers an effective framework to inject explicit actions into agent decision, toward reliable audio intelligence.

Limitations

Despite promising results, our study presents several limitations that offer avenues for future exploration.

Token imbalance and rewrite sparsity. Our training data exhibits an inherent imbalance across action tokens (<internal>, <external>, <rewrite>). While both <internal> and <external> achieve high F1 scores, <rewrite> remains under-trained on many datasets. This sparsity partly reflects that certain audio QA datasets rarely require rewriting but this contextual information sparsity also reveals a modeling challenge. Future work may explore more principled strategies for balancing token distribution or adaptively reshaping decision boundaries, especially under varying persona settings or task configurations.

Limited ASR training subset. Our current ASR experiments are trained on a restricted subset of data. While the model already achieves strong performance, it likely underutilizes the available signal. Scaling up training with larger ASR datasets or augmenting with synthetic audio variants may unlock further gains.

No exploration of transfer or multi-external setups. We do not yet study transfer capabilities. For example, training with one external ASR model and testing with another. Moreover, our current system only accepts a single external prediction. Extending the framework to handle multiple external models, each represented by distinct decision tokens, could significantly improve robustness and enable broader deployment across diverse real-world pipeline

Acknowledgments

This paper is also partially supported by Research and Development Center for Large Language Models, National Institute of Informatics (NII LLMC).

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. 2019. A multi-room reverberant dataset for sound event localization and detection. *Preprint*, arXiv:1905.08546.

A. Calcus. 2024. Development of auditory scene analysis: a mini-review. *Frontiers in Human Neuroscience*, 18:1352247.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything ami meet-

ing corpus. *Language Resources and Evaluation*, 41(2):181–190.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*, interspeech_2021. ISCA.

Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, and Zhou Zhao. 2025. Wavrag: Audio-integrated retrieval augmented generation for spoken dialogue models. *Preprint*, arXiv:2502.14727.

Kanzhi Cheng, Yantao Li, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. 2024. Vision-language models can self-improve reasoning via reflection. *arXiv preprint arXiv:2411.00855*.

Yangui Fang, Baixu Cheng, Jing Peng, Xu Li, Yu Xi, Chengwei Zhang, and Guohui Zhong. 2025. Fewer hallucinations, more verification: A three-stage llm-based framework for asr error correction. *arXiv preprint arXiv:2505.24347*.

B. Galantucci, C. A. Fowler, and M. T. Turvey. 2006. The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, 13(3):361–377. Erratum in: Psychon Bull Rev. 2006 Aug;13(4):742.

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

Arushi Goel, Karan Sapra, Matthieu Le, Rafael Valle, Andrew Tao, and Bryan Catanzaro. 2024. Omcat: Omni context aware transformer. *arXiv preprint arXiv:2410.12109*.

Robie Gonzales and Frank Rudzicz. 2024. A retrieval augmented approach for text-to-music generation. In *Proceedings of the 3rd Workshop on NLP for Music and Audio (NLP4MusA)*, pages 31–36, Oakland, USA. Association for Computational Lingustics.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. *TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation*, page 198–208. Springer International Publishing.

Rui Hu, Delai Qiu, Shuyu Wei, Jiaming Zhang, Yining Wang, Shengping Liu, and Jitao Sang. 2025. Investigating and enhancing vision-audio capability in omnimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7452–7463, Vienna, Austria. Association for Computational Linguistics.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *Preprint*, arXiv:2301.12661.

M. Kaiser, D. Senkowski, and J. Keil. 2021. Mediofrontal theta-band oscillations reflect top-down influence in the ventriloquist illusion. *Human Brain Mapping*, 42(2):452–466. Epub 2020 Oct 14.

Yuma Koizumi, Yasunori Ohishi, Daisuke Niizumi, Daiki Takeuchi, and Masahiro Yasuda. 2020. Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval. *Preprint*, arXiv:2012.07331.

C. Lebiere and J. R. Anderson. 2011. Cognitive Constraints on Decision Making under Uncertainty. *Frontiers in Psychology*, 2:305.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Xingwei Sun, Tianzi Wang, Junbo Zhang, and Jian Luan. 2025a. Miaqa submission for dcase 2025 challenge task 5: A reinforcement learning driven audio question answering method. Technical report, DCASE2025 Challenge.

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

Yen-Ting Lin, Zhehuai Chen, Piotr Żelasko, Zhen Wan, Xuesong Yang, Zih-Ching Chen, Krishna C Puvvada, Ke Hu, Szu-Wei Fu, Jun Wei Chiu, et al. 2025. Neko: Cross-modality post-recognition error correction with tasks-guided mixture-of-experts language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 222–236.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. 2025. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Christian Marinoni, Riccardo Fosco Gramaccioni, Changan Chen, Aurelio Uncini, and Danilo Comminiello. 2024. Overview of the l3das23 challenge on audio-visual extended reality. *Preprint*, arXiv:2402.09245.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav

Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Thomas O. Nelson. 1990. Metamemory: A Theoretical Framework and New Findings. In Gordon H. Bower, editor, *Psychology of Learning and Motivation*, volume 26, pages 125–173. Academic Press.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D. Shulman, Boris Ginsburg, Shinji Watanabe, and Georg Kucsko. 2021. Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. In *Interspeech 2021*, pages 1434–1438.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5206–5210.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

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

Matthew Renze and Erhan Guven. 2024. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024a. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024b. Mmau: A massive multi-task audio understanding and reasoning benchmark. *Preprint*, arXiv:2410.19168.

Laela Sayigh, Mary Ann Daher, Julie Allen, Helen Gordon, Katherine Joyce, Claire Stuhlmann, and Peter Tyack. 2016. The watkins marine mammal sound database: an online, freely accessible resource. In *Proceedings of Meetings on Acoustics*, volume 27, page 040013. Acoustical Society of America.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025. Canary-1b-v2 & parakeet-tdt-0.6b-v3: Efficient and high-performance models for multilingual asr and ast. *Preprint*, arXiv:2509.14128.

Robert L Selman and Diane F Byrne. 1974. A structural-developmental analysis of levels of role taking in middle childhood. *Child development*, pages 803–806.

Zihan Song, Xin Wang, Zi Qian, Hong Chen, Longtao Huang, Hui Xue, and Wenwu Zhu. Modularized self-reflected video reasoner for multimodal llm with application to video question answering. In *Forty-second International Conference on Machine Learning*.

Zhen Wan, Chao-Han Huck Yang, Yahan Yu, Jinchuan Tian, Sheng Li, Ke Hu, Zhehuai Chen, Shinji Watanabe, Fei Cheng, Chenhui Chu, et al. 2025. Speechiq: Speech-agentic intelligence quotient across cognitive levels in voice understanding by large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30381–30398.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Zixuan Wang, Chi-Keung Tang, and Yu-Wing Tai. 2025. Audio-agent: Leveraging llms for audio generation, editing and composition. *Preprint*, arXiv:2410.03335.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *Preprint*, arXiv:2408.16725.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

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

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

Chao-Han Huck Yang, Sreyan Ghosh, Qing Wang, Jaeyeon Kim, Hengyi Hong, Sonal Kumar, Guirui Zhong, Zhifeng Kong, S Sakshi, Vaibhavi Lokegaonkar, Oriol Nieto, Ramani Duraiswami, Dinesh Manocha, Gunhee Kim, Jun Du, Rafael Valle, and Bryan Catanzaro. 2025a. Multi-domain audio question answering toward acoustic content reasoning in the dcase 2025 challenge. *Preprint*, arXiv:2505.07365.

Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. 2023a. Generative speech recognition error correction with large language models and task-activating prompting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. 2025b. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023b. Mm-react: Prompting chatgpt for multimodal reasoning and action. *Preprint*, arXiv:2303.11381.

Lu Zhang, Tiancheng Zhao, Heting Ying, Yibo Ma, and Kyusong Lee. 2024. Omagent: A multi-modal agent framework for complex video understanding with task divide-and-conquer. *Preprint*, arXiv:2406.16620.

Jinming Zhao, Gholamreza Haffari, and Ehsan Shareghi. 2023. Generating synthetic speech from SpokenVocab for speech translation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1975–1981, Dubrovnik, Croatia. Association for Computational Linguistics.

A Failure Case of Multimodal GER

A.1 Amplifying Hallucination

This is often a situation where external audio perception is not strong enough or N-best decoding introduces noise for GER. Especially when the omni model is encouraged to take all hypothesses in consideration.

ASR Failure Case

Internal: insane
External (Whisper 5-best): [' Gimseeinnnnnn', ' You can say.', ' Insta.', ' Wednesday.', " I'm from Phelps County, I'm gonna see what this guy's doing."]

GER Pred: I'm from Phelps County, I'm gonna see what this guy's doing
Gold: insane

A.2 Overcorrection

Due the language modeling nature, LLMs tend to revise the transcription to be more likea complete sentence, which sometimes caused "overcorrection."

ASR Failure Case

Internal: you in the way marguerite but how
External (Whisper 5-best): [' you ll in the way marguerite but how', 'you in the way marguerite but how', 'you in the way marguerite but how.', ' you in the way marguerite but how.', " you in the way marguerite but how."]

GER Pred: you are in the way marguerite but how
Gold: you in the way marguerite but how

A.3 Prompt Ablations in Preliminary SFT

We conduct these experiments during early-stage exploration: for inputs we investigated both (a) audio + external whisper 5-best and (b) audio + internal 1-best + external whisper 5-best, together with four prompting strategies instructing the model to emphasize internal hypotheses, external hypotheses, audio grounding, or a balanced fusion. The average WER results on OpenASR are shown in Table 1.

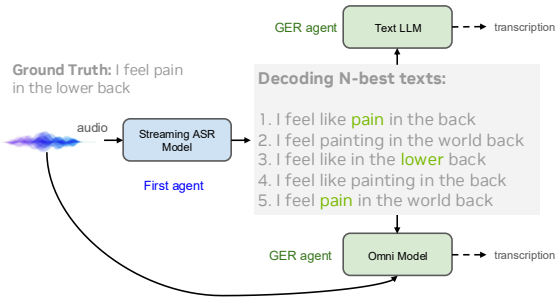


Figure 5: Text-base GER uses only ASR hypotheses. This setup fails to correct deletion or hallucination if all hypotheses are wrong. Multimodal GER include the original audio as grounding to improve error correction.

A.4 Zero-shot Qwen in Preliminary

We test three different zero-shot prompting strategies as shown in Table 2. We find that zero-shot decisions are highly prompt-sensitive: the model often collapses into trivial heuristics, while the balanced prompt produces unstable and inconsistent behavior across samples. The corresponding 2×2 confusion matrices show large off-diagonal mass for all zero-shot prompts, confirming that zero-shot Qwen does not perform genuine self-reflection.

B Prompt Template

Here is the prompt for AudioQA, we explicitly prompt the model to ourput `<external>` only when the internal prediction is wrong while external perception is correct.

```
You are an audio understanding model
with access to three inputs:
(1) The original audio;
(2) One answer candidate generated by
another model (external);
(3) Your own prediction (internal).


Your task is to decide which of the
following strategies to apply:
- If your internal prediction is correct
and acceptable, output <internal> and
repeat your answer.
- If the external candidate is correct
while your internal prediction is
incorrect, output <external> and use
the external answer.
- If all given answers are incorrect,
output <rewrite> and re-answer the
question correctly based only on the
original audio.


Return the selected token
(<internal>/<external>/<rewrite>)
followed by your final answer.
```

C Additional Agentic Studies on ASR and AudioQA Tool Calling

C.1 Active Perception via DSP Tool Calling

While the main *Speech-Hands* framework arbitrates between internal and external semantic hypotheses, our error analysis (Section 6.3) reveals that "Dual-Failure" modes often stem from intrinsic audio degradation. For example, background noise in Soundscape QA and poor microphone quality in AMI meetings are issues that could be mitigated through additional post-processing.

To address this, we conduct an extra mini-experiment expanding the agent's action space from purely generative decisions to **Active Perception**. We integrate open license Digital Signal Processing (DSP) tools as executable actions, allowing the model to "clean its ears" before transcription.

C.2 Tool-Based Action Space

We extend the action token set $\mathcal{A} = \{$`<internal>`, `<external>`, `<rewrite>`$\}$ with two signal enhancement tools:

Background Noise Removal (BNR) (`<tool:bnr>`). This action triggers the open **NVIDIA BNR inference microservices**, designed to suppress non-speech broad-spectrum noise (e.g., traffic, sirens) while preserving emotive vocal tones.

- **Trigger Logic:** The agent emits this token when the internal audio embedding suggests high background entropy or when the initial decode contains disfluency markers like `[noise]` or `[unintelligible]`.

Studio Voice Restoration (`<tool:studio>`). This action triggers **NVIDIA Studio Voice** (via Maxine SDK), which uses generative upsampling to reconstruct high-frequency harmonics (48kHz upsampling) from low-bandwidth input.

- **Trigger Logic:** Activated when the agent detects low sample rates ($<$ 16kHz), narrowband telephony audio, or significant room reverberation (common in the AMI corpus).

C.3 Experimental Results

We evaluate this *Active Speech-Hands* setup on the two most challenging subsets identified in our main results: **AMI** (noisy meetings) and **Soundscape QA** (environmental noise). As shown in Table 7, enabling DSP tools significantly benefits these noise-sensitive tasks.

Metric / Dataset	Standard *(Passive)*	w/ Active Tools *(BNR + Studio)*
Speech Recognition (WER ↓)		
AMI (Meeting)	15.03	**13.41** (-1.62)
VoxPopuli (Reverb)	6.49	**5.82** (-0.67)
Audio Reasoning (Acc % ↑)		
Soundscape QA	59.40	**63.15** (+3.75)
Complex QA	85.70	**86.44** (+0.74)

Table 7: Mini-experiment results on Active Perception. Integrating `<tool:bnr>` and `<tool:studio>` improves performance on noisy datasets (AMI) and acoustic reasoning tasks compared to the passive Speech-Hands baseline.

Analysis. On the AMI dataset, `<tool:studio>` successfully reconstructs fricatives lost in distant-microphone recordings, reducing WER by 1.62%. In Soundscape QA, `<tool:bnr>` helps segregate foreground events from background clutter.

Active Tool Case Study (Soundscape)

Audio: A faint voice speaking over loud construction drilling noise.
Q. What is the speaker asking for?

Standard (Passive): *Internal:* [Noise]
External: Help me with the grill.
Prediction: `<external>` Help me with the grill. (✗)

Active (Ours): *Action:* `<tool:bnr>` → *Cleaned Audio*
New Internal: Help me with the drill.
Prediction: `<internal>` Help me with the drill. (✓)

This demonstrates that for highly degraded inputs, agentic reasoning should precede perception, which decides *how* to listen is as important as deciding *what* was heard.

D Audio Dataset Details

D.1 Dataset Details for Speech Recognition

To ensure a fair and balanced evaluation across diverse speech corpora, we uniformly sampled up to 20,000 audio-question pairs from each dataset for training as in Table 8. All datasets were aligned to a consistent prompt-question-answer format to support unified multi-dataset training.

Dataset	AMI	Tedlium	gigaspeech	spgispeech	Voxpopuli	Libri	Libri-clean	Libri-other
Sampling #	Subset in Speech-Hands Training							
Train	20,000	20,000	9,389	20,000	20,000	20,000	20,000	20,000

Table 8: Number of training samples used from each dataset. For GigaSpeech, only 9,389 valid samples met our filtering criteria.

D.2 Dataset Details for DCASE2025 AudioQA

Subset	#Train / #Dev	<in>/<ex>/<re>
Bio-acoustic QA	0.7K / 0.2K	338/234/168
Soundscape QA	1.0K / 0.6K	604/182/252
Complex QA	6.4K / 1.6K	4,267/1,785/391

Table 9: Statistics of the three DCASE2025 AudioQA subsets used in our experiments. **<in>/<ex>/<re>** shows the token distribution in training w/ majority sampling.

The DCASE2025 AudioQA benchmark comprises three complementary multiple-choice question-answering subsets, each designed to evaluate a different aspect of audio reasoning.

D.3 Bioacoustics QA

This subset targets perceptual and cognitive grounding in marine bioacoustics. It includes questions about 31 marine mammal species with diverse acoustic ranges, habitats, and vocalization durations. Tasks include species classification, vocalization type recognition, factual retrieval, interpretation of acoustic features, and comparative reasoning. The dataset includes approximately 0.7K training and 0.2K development QA pairs. Audio clips range in sample rate from 600 Hz to 160 kHz and in duration from 0.4 s to 625 s, allowing evaluation under highly varied acoustic conditions. All audio is sourced from the Watkins Marine Mammal Sound Database (Woods Hole Oceanographic Institution; New Bedford Whaling Museum), and usage of audio beyond the provided splits is strictly prohibited.

D.4 Temporal Soundscapes QA

This subset focuses on temporal reasoning over sound events, encompassing 26 event classes. Questions require identifying active sound classes, temporal ordering, timestamp estimation (onset, offset, duration), and event comparison. The subset comprises approximately 1.0K training and 0.6K development QA pairs. Audio clips are mono-channel, 10 seconds long, and sampled at 32–48 kHz. Most clips correspond to a single QA item, while a small portion supports multiple questions. All annotations are manually verified, which include event types, timestamps, and answers. Audio is sourced from NIGENS, L3DAS23 Challenge (Marinoni et al., 2024), and TAU Spatial Sound Events 2019 datasets (Adavanne et al., 2019).

D.5 Complex QA (MMAU)

This subset evaluates high-level reasoning over natural sound scenes. Each instance consists of a 10-second, 16 kHz audio clip paired with a question requiring reasoning over acoustic, temporal, and contextual cues (*i.e.*, overlapping events, implied sequences, or abstract relationships.) Tasks are inspired by the MMAU Sound benchmark (Sakshi et al., 2024b) and extend its principles to more diverse soundscapes. The data include approximately 6.4K training and 1.6K development QA pairs. Audio clips are sourced from AudioSet and Mira.

E Cases On ASR

We provide several examples to show the effectiveness of our Speech-Hands in ASR tasks.

ASR Case 1

Q. What is the correct transcription of this audio?

Internal: water and fresh stalls to take on there
External: we've water and fresh stores to tick on there
Rewrite: we've water and fresh stores to take on there

Pred: `<rewrite>` we've water and fresh stores to take on there
Gold: we've water and fresh stores to take on there

ASR Case 2

Q. What is the correct transcription of this audio?

Internal: then has nothing happened with a voice like a whip mr wicker said start at once
External: then as nothing happened with a voice like a whip mister wigger said start at once
Rewrite: then as nothing happened with a voice like a whip mister wuthers said start at once

Pred: <external> then as nothing happened with a voice like a whip mister wigger said start at once
Gold: then as nothing happened with a voice like a whip mister wicker said start at once

F Cases On AudioQA

For each token we pick up 5 examples to illustrate how action tokens influence the final correctness. For <rewrite> token, it works not only when the internal prediction is conflicted with the external prediction, but also show robust and accurate decision when hypothesses generate the same but wrong prediction. This is extremely important when facing challenging tasks, the omni-agent should be confident to refuse wrong hypothesses. Also for <internal> and <external> tokens that have more frequencies, they should be accurately distinguished when conflicts, and action tokens complete a good job on recognizing the true answer.

AudioQA Case 1

Q. Based on the audio, which natural phenomenon could be occurring?
Options:
A. Earthquake
B. Thunderstorm
C. Forest fire
D. Snowstorm

Internal pred: B. Thunderstorm
External pred: B. Thunderstorm

Pred: <rewrite> C. Forest fire
Gold: <rewrite> C. Forest fire

AudioQA Case 2

Q. What activity is likely occurring based on the described sounds?
Options:
A. Driving a car
B. Planting a garden
C. Painting a picture
D. Cooking a meal

Internal pred: A. Driving a car
External pred: B. Planting a garden

Pred: <rewrite> D. Cooking a meal
Gold: <rewrite> D. Cooking a meal

AudioQA Case 3

Q. Based on the audio, what is the most likely environment in which the sounds are occurring?
Options:
A. An indoor setting
B. A rural countryside
C. A busy city street
D. A quiet residential neighborhood

Internal pred: A. An indoor setting
External pred: B. A rural countryside

Pred: <rewrite> C. A busy city street
Gold: <rewrite> C. A busy city street

AudioQA Case 4

Q. What background sound is present in the audio clip?
Options:
A. Waves crashing
B. Birds chirping
C. Bells ringing
D. Running car engine

Internal pred: A. Waves crashing
External pred: D. Running car engine

Pred: <rewrite> C. Bells ringing
Gold: <rewrite> C. Bells ringing

AudioQA Case 5

Q. What might the purpose of tapping the metal object in the background be?
Options:
A. To emphasize the speaker's instructions
B. To distract from the speaker's sad tone
C. To demonstrate its material quality
D. To create a rhythmic background

Internal pred: A. To emphasize the speaker's instructions
External pred: A. To emphasize the speaker's instructions

Pred: <rewrite> C. To demonstrate its material quality
Gold: <rewrite> C. To demonstrate its material quality

AudioQA Case 6

Q. What type of mood is conveyed through the musical elements in this audio?
Options:
A. Calm and soothing
B. Angry and aggressive
C. Joyful and uplifting
D. Sad and reflective

Internal pred: B. Angry and aggressive
External pred: D. Sad and reflective

Pred: <external> D. Sad and reflective
Gold: <external> D. Sad and reflective

AudioQA Case 7

Q. Based on the audio description, what is likely happening in the background?
Options:
A. A calm evening
B. A windy day
C. An earthquake
D. A quiet morning

Internal pred: B. A windy day
External pred: C. An earthquake

Pred: <external> C. An earthquake
Gold: <external> C. An earthquake

AudioQA Case 8

Q. What type of sound is present in the background of the audio clip?
Options:
A. Car engine
B. Ocean waves
C. Upbeat synthesized music
D. Bird chirping

Internal pred: B. Ocean waves
External pred: C. Upbeat synthesized music

Pred: <external> C. Upbeat synthesized music
Gold: <external> C. Upbeat synthesized music

AudioQA Case 9

Q. Based on the audio description, what is the primary focus of the sounds?
Options:
A. A quiet library setting
B. A peaceful nature scene
C. A busy city street
D. A defense attack scenario

Internal pred: A. A quiet library setting
External pred: D. A defense attack scenario

Pred: <external> D. A defense attack scenario
Gold: <external> D. A defense attack scenario

AudioQA Case 10

Q. Based on the audio description, what type of activity is most likely taking place?
Options:
A. Gardening
B. Woodworking
C. Lock-picking
D. Cooking

Internal pred: B. Woodworking
External pred: C. Lock-picking

Pred: <external> C. Lock-picking
Gold: <external> C. Lock-picking

AudioQA Case 11

Q. What element in the audio contributes to the emotional depth of the song besides the vocals?
Options:
A. The language spoken
B. The steady drum beats
C. The groovy bass line
D. The keyboard harmony

Internal pred: D. The keyboard harmony
External pred: C. The groovy bass line

Pred: <internal> D. The keyboard harmony
Gold: <internal> D. The keyboard harmony

AudioQA Case 12

Q. Based on the audio description, what type of environment is suggested by the background sounds?
Options:
A. A serene forest
B. A quiet library
C. A beach with waves
D. A busy city street

Internal pred: D. A busy city street
External pred: D. A busy city street

Pred: <internal> D. A busy city street
Gold: <internal> D. A busy city street

AudioQA Case 13

Q. What might the purpose of the whistle and crinkling leaves in the background be?
Options:
A. To create a suspenseful atmosphere
B. To signal the attention of someone nearby
C. To mimic a bustling city environment
D. To indicate the presence of wildlife

Internal pred: B. To signal the attention of someone nearby
External pred: B

Pred: <internal> B. To signal the attention of someone nearby
Gold: <internal> B. To signal the attention of someone nearby

AudioQA Case 14

Q. Why does the conversation feature electronic beats and rhythmic cymbal sounds?
Options:
A. To mimic the sounds of a busy environment
B. To create a sense of urgency in the interaction
C. To drown out background noise
D. To enhance the emotional depth of the speech

Internal pred: B. To create a sense of urgency in the interaction
External pred: D. To enhance the emotional depth of the speech

Pred: <internal> B. To create a sense of urgency in the interaction
Gold: <internal> B. To create a sense of urgency in the interaction

AudioQA Case 15

Q. What sound appears earliest in the audio?
Options:
A. Ticking
B. Accelerating, revving, vroom
C. Idling
D. Car

Internal pred: C. Idling
External pred: D. Car

Pred: <internal> C. Idling
Gold: <internal> C. Idling