Focus Then Listen: An Empirical Study of Plug-and-Play Audio Enhancer for Noise-Robust Large Audio Language Models

Han Yin [1] **Yang Xiao** [2] **Younghoo Kwon** [1] **Ting Dang** [2] **Jung-Woo Choi** [1]

Abstract

Large audio language models (LALMs) are a class of foundation models for audio understanding. Existing LALMs tend to degrade significantly in real-world noisy acoustic conditions where speech and non-speech sounds interfere. While noise-aware fine-tuning can improve robustness, it requires task-specific noisy data and expensive retraining, limiting scalability. To address this issue, we propose Focus-Then-Listen (FTL), a plug-and-play audio enhancer that improves LALMs' noise robustness. Specifically, FTL first separates the input waveform into speech and non-speech, and a modality router is applied to predict the target audio modality (e.g., speech) based on the user's instruction. Finally, a modality-aware fusion block generates a task-adaptive enhanced signal for improved downstream perception and reasoning. Experiments across multiple LALMs and tasks show that FTL improves performance across different noise levels without fine-tuning on LALMs.

1. Introduction

Large audio language models (LALMs) have recently emerged as a powerful paradigm for unified audio understanding and reasoning (Chu et al., 2023; Ghosh et al., 2024; Xie et al., 2025). By integrating audio perception with large language models (LLMs), LALMs enable a wide range of applications, including speech recognition, acoustic scene analysis, and audio question answering (Yu et al., 2024; Deshmukh et al., 2023; Wang et al., 2025).

Noise robustness remains a fundamental challenge for LALMs (Li et al., 2026). Here, the noise refers to acoustic signals that are irrelevant to the user's intent in a given task.

[1]School of Electrical Engineering, KAIST, Daejeon, Republic of Korea [2]University of Melbourne, Australia. Correspondence to: Jung-Woo Choi <jwoo@kaist.ac.kr>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

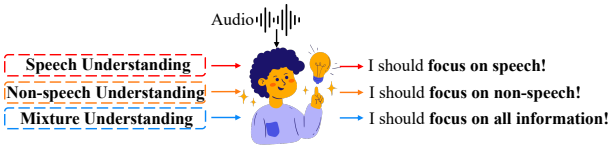

Figure 1. Process of human audio understanding.

For instance, in speech understanding tasks, non-speech sounds can be the noise, whereas in environmental sound analysis, speech may act as interference. In real-world environments, audio inputs are rarely clean and often contain multiple overlapping or irrelevant components. Without sufficient robustness to such task-irrelevant signals, LALMs may misinterpret the user's intent, resulting in degraded interaction quality and unreliable system behavior, particularly in safety-critical applications (Torad et al., 2022; Schmidt et al., 2019; Zhang et al., 2015).

Recent work has begun to investigate this problem. SSEU-Bench (Yin & Choi, 2026) explicitly models the coexistence of speech and non-speech sounds and considers their energy imbalance across diverse scenarios. An important observation is that cross-component interference significantly affects model performance: when performing speech understanding, strong non-speech sounds can degrade recognition, and similarly, dominant speech can negatively impact non-speech sound understanding. To address this issue, SSEU-Bench uses chain-of-thought (CoT) prompting to decompose complex audio understanding into simpler steps. However, the improvement is mainly observed in audio tagging tasks, and CoT often requires task-specific prompt design. Another straightforward approach to enhance robustness is noise-aware training, which involves fine-tuning models on large-scale datasets augmented with various noise types (Hu et al., 2024; Ding et al., 2025). This paradigm requires extensive data curation, as covering the infinite variability of real-world noise is practically infeasible. In addition, fine-tuning may also lead to catastrophic forgetting or degrade performance on clean data (Luo et al., 2025; Zhai et al., 2023; Yin et al., 2015). In SEE (Zhang et al., 2026), researchers propose an embedding-based approach for developing noise-robust LALMs, but assumes that noise is explicitly pre-defined (e.g., Gaussian noise) and the isolated pure-noise recordings are required during training. This assumption is incompatible with our setting, where noise

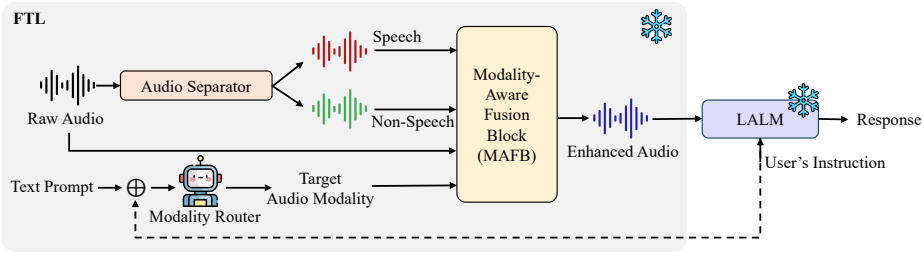

Figure 2. Overview of proposed audio enhancer (FTL) for noise-robust large audio language models.

cannot be pre-defined but is task-dependent: non-speech acts as noise for speech tasks, and vice versa.

To address these issues, we explore to use an audio enhancer, i.e., Focus-Then-Listen (FTL), to improve LALMs' noise robustness. Our motivation stems from the human audio understanding process. As illustrated in Fig. 1, when confronted with audio, humans selectively focus on the component relevant to their intent. Inspired by this, FTL infers the task-relevant audio modality from the user's instruction and produces a filtered, modality-aligned signal for the LALM, which improves downstream perception and reasoning in noisy conditions. Our contributions are as follows:

- We propose FTL, a plug-and-play audio enhancer that improves LALMs' noise robustness. To the best of our knowledge, this is the first work to explore mitigating speech and non-speech interference for LALMs via instruction-aware audio enhancement.
- Experiments across multiple LALMs and benchmarks demonstrate the effectiveness of the proposed FTL in both audio perception and reasoning tasks[1].

2. Focus Then Listen

2.1. Overview

As shown in Fig. 2, in FTL, we first use an audio separator to separate the raw input audio into speech and non-speech tracks, which can be expressed as:

$$\mathbf{S_{sp}}, \mathbf{S_{ns}} = Sep(\mathbf{S_{ra}}) \tag{1}$$

where $Sep(\cdot)$ is the audio separator, $\mathbf{S_{ra}}$ is the raw audio, $\mathbf{S_{sp}}$ and $\mathbf{S_{ns}}$ are the separated speech and non-speech, respectively. We then introduce a modality router to determine the target audio modality m based on the user's instruction. If the task only requires speech-related information, the router should output "*speech*"; if the task focuses on non-speech content, the router outputs "*non-speech*". For more complex tasks that require both modalities, the router outputs "*mixture*". Specifically, we use an LLM as the modality router; the detailed prompt used for the LLM is provided on Appendix A.1. Finally, we employ a modality-aware fusion block (MAFB) to generate task-adaptive enhanced audio conditioned on the selected modality. The goal of the MAFB is to refine the acoustic signal to better align with the user's task. Through FTL, we aim to amplify task-relevant information while suppressing irrelevant components in the audio, allowing the downstream LALM to focus more effectively on informative acoustic cues.

2.2. Modality-Aware Fusion Block

The MAFB is designed to generate task-adaptive enhanced audio based on the modality selected by the router. Specifically, the enhanced audio $\mathbf{S_{en}}$ is computed as:

$$\mathbf{S_{en}} = \begin{cases} \alpha_{sp}\mathbf{S_{sp}} + (1-\alpha_{sp})\mathbf{S_{ra}}, & \text{if } m = \text{“}\textit{speech}\text{”} \\ \alpha_{ns}\mathbf{S_{ns}} + (1-\alpha_{ns})\mathbf{S_{ra}}, & \text{if } m = \text{“}\textit{non-speech}\text{”} \\ \mathbf{S_{ra}}, & \text{if } m = \text{“}\textit{mixture}\text{”} \end{cases} \tag{2}$$

where m denotes the target audio modality predicted by the router. The coefficients α_{sp} and α_{ns} are hyperparameters that control the degree of enhancement (ranging from 0 to 1). Specifically, we set $\alpha_{sp} = 0.5$ and $\alpha_{ns} = 0.9$, which are determined empirically through ablation studies. Detailed analyses are provided in Appendix B.

The MAFB performs modality-aware signal fusion between separated signals and raw audio. This design balances modality enhancement and signal fidelity. When separated signals contain artifacts, mixing in original audio preserves natural acoustics and improves LALMs' robustness.

2.3. Audio Separator

In FTL, we employ an audio separation model to separate speech from non-speech components. To achieve this, we first consider existing state-of-the-art (SOTA) pre-trained models: SE-Mamba (**SEM**) (Chao et al., 2024) and SAM-Audio (**SAM**) (Shi et al., 2025).

Specifically, SEM is a GAN (Goodfellow et al., 2014)-based speech enhancement model; the enhanced speech is first estimated from the mixture, and the non-speech signal is obtained by subtracting the enhanced speech from the mixture. SAM is a generative separation model that simultaneously estimates both the target and residual stems from an audio mixture, conditioned on text or visual prompts; we use a text prompt with the content "*speech*".

[1] https://sites.google.com/view/ftl-lalm

However, SEM is trained with speech enhancement objectives instead of separation, and SAM may generate signal components not present in the raw audio, which can potentially mislead downstream audio understanding tasks. Therefore, we develop **SNSep**, a separator specialized for speech and non-speech separation, which operates in the short-time Fourier transform domain using a masking-based approach. Specifically, we adopt the separation network from AudioSep (Liu et al., 2024; Yin et al., 2024) as the backbone. SNSep has a dual-decoder architecture: one decoder reconstructs the speech track, while a parallel decoder independently extracts the non-speech track.

3. Experimental Setups

3.1. Detailed Configurations

SNSep Training: For training SNSep, we sample 50 hours of speech from LJSpeech, Librispeech, VoxPopuli, and GigaSpeech training sets (Ito & Johnson, 2017; Panayotov et al., 2015; Wang et al., 2021; Chen et al., 2021) and 50 hours of non-speech from VocalSound, VGGSound, CochlScene, AudioSet, FSD50K, and UrbanSound8K training sets (Gong et al., 2022; Chen et al., 2020; Jeong & Park, 2022; Gemmeke et al., 2017; Fonseca et al., 2021; Salamon et al., 2014). During training, a speech and a non-speech sample are mixed with an SNR randomly selected from -10 dB to 10 dB as the input, and all audio samples are resampled to 16 kHz. For other configurations, we follow the previous work (Liu et al., 2024).

Modality Router and LALM: For the modality router, we use Qwen3-8B (Yang et al., 2025) and ChatGPT5.2. For the LALM, we adopt Audio Flamingo 3 (**AF3**) (Goel et al., 2025), Fun-Audio-Chat (**FAC**) (Team et al., 2025), and Qwen3-Omni (**Q3O**) (Xu et al., 2025).

3.2. Evaluation

We investigate the effectiveness of FTL on **audio perception** and **audio reasoning** tasks. Audio perception tasks primarily assess the model's ability to recognize and understand low-level acoustic events and spoken content, while audio reasoning tasks require higher-level semantic inference and compositional reasoning over auditory objects. Through these experiments, we aim to investigate whether FTL consistently improves both perceptual understanding and reasoning capabilities in LALMs.

Audio Perception: We use SSEU-Bench to evaluate audio perception performance in noisy conditions, where each audio sample is mixed with a speech and a non-speech sound with a specific SNR. Two classic tasks in speech and environmental sound domains are included : Automatic Speech Recognition (ASR) and Audio Tagging (AT). For ASR, the LALM is required to output the spoken content of the speaker. For AT, we require the LALM to detect non-speech sound events within the audio. Detailed instructions used for the two tasks are provided in Appendix A.2.

Table 1. ASR performance of LALMs on SSEU-Bench. Metric is WER(%). "SNR-Speech" refers to speech to non-speech ratio, where "SNR-Speech=+∞" means pure speech.

LALM	FTL	SNR-Speech (dB)					
		+∞	10	5	0	-5	-10
AF3	✗	2.18	2.71	3.27	4.73	10.45	27.45
	✔	**2.17**	**2.66**	**3.20**	**4.61**	**9.83**	**25.39**
FAC	✗	**2.61**	3.38	3.99	5.75	12.54	31.67
	✔	**2.61**	**3.24**	**3.86**	**5.44**	**11.54**	**28.41**
Q3O	✗	**2.16**	**2.31**	2.56	3.64	7.04	20.42
	✔	2.23	**2.31**	**2.49**	**3.38**	**5.97**	**18.61**

Table 2. AT performance of LALMs on SSEU-Bench. Metric is mAP(%). "SNR-Non-Speech" refers to non-speech to speech ratio, where "SNR-Non-Speech=+∞" means pure non-speech.

LALM	FTL	SNR-Non-Speech (dB)					
		+∞	10	5	0	-5	-10
AF3	✗	38.80	36.18	34.56	31.00	28.86	27.36
	✔	**39.28**	**39.26**	**38.95**	**38.19**	**34.94**	**31.56**
FAC	✗	36.34	21.27	18.33	17.54	16.98	16.34
	✔	**36.64**	**31.97**	**30.32**	**27.88**	**24.89**	**20.75**
Q3O	✗	**44.43**	39.75	38.12	34.84	33.20	31.33
	✔	44.27	**43.48**	**42.25**	**40.32**	**39.20**	**37.27**

Audio Reasoning: MMAU-Pro (Kumar et al., 2026) is a widely used audio reasoning benchmark, which comprises various audio-based question–answer (QA) pairs. However, MMAU-Pro does not provide specific SNRs for speech and non-speech components within an audio sample. Therefore, we curate a new subset of MMAU-Pro with controllable SNR conditions, i.e., **MMAU-Pro-Ctrl**.

Specifically, we manually collect 130 speech- and 130 non-speech-QAs from MMAU-Pro. For the speech-QA subset, we ensure that the audio consists of clean speech (4s to 300s) with questions explicitly target speech content. Conversely, the non-speech subset contains non-speech audio (5s to 293s) with questions. To simulate realistic noisy speech-QAs, we utilize the non-speech samples as the noise. For each pair, noise shorter than the speech is randomly inserted, whereas longer noise is cropped to match its duration. The same mixing protocol is used for non-speech QAs, with speech treated as noise. Following SSEU-Bench, we range the SNR from 10 dB to -10 dB.

Metrics: We use **Word Error Rate (WER)** to evaluate ASR. For AT, we use **mean Average Precision (mAP)** for evaluation. For reasoning tasks, we follow MMAU-Pro and use the averaged accuracy for evaluation, denoted as **QA-ACC**. In addition, we report the **Correct Rate (CR)** to measure the performance of the modality router, which is defined as the proportion of samples where the target audio modality is correctly predicted.

Table 3. Reasoning performance (QA-ACC(%)) on MMAU-Pro-Ctrl with different modality routers (LLM: Qwen3-Omni).

Speech Reasoning								
FTL	Modality Router	CR(%)	SNR-Speech (dB) $+\infty$	10	5	0	-5	-10
✗	-	-	**75.4**	75.4	**75.4**	74.6	73.1	70.0
✔	Qwen3-8B	23.8	**75.4**	74.6	74.6	73.8	**74.6**	70.0
✔	ChatGPT5.2	88.5	**75.4**	**76.2**	**75.4**	**75.4**	**74.6**	**73.1**
✔	GroundTruth	100.0	76.2	75.4	75.4	75.4	73.8	72.3
Non-Speech Reasoning								
FTL	Modality Router	CR(%)	SNR-Non-Speech (dB) $+\infty$	10	5	0	-5	-10
✗	-	-	**43.1**	35.4	38.5	37.7	36.2	34.6
✔	Qwen3-8B	0.0	**43.1**	35.4	38.5	37.7	36.2	34.6
✔	ChatGPT5.2	47.7	41.5	**42.3**	**43.1**	**40.0**	**39.2**	**38.5**
✔	GroundTruth	100.0	42.3	42.3	40.8	40.0	39.2	38.5

Table 4. Performance of Audio Flamingo 3 with different audio separators in FTL on SSEU-Bench.

ASR Performance, Metric: WER (%)						
FTL	Sep	SNR-Speech (dB) 10	5	0	-5	-10
✗	-	2.71	3.27	4.73	10.45	27.45
✔	SAM	2.83	3.31	4.93	10.40	28.72
✔	SEM	**2.62**	**3.03**	**4.08**	**8.07**	**23.83**
✔	SNSep	2.66	3.20	4.61	9.83	25.39
AT Performance, Metric: mAP (%)						
FTL	Sep	SNR-Non-Speech (dB) 10	5	0	-5	-10
✗	-	36.18	34.56	31.00	28.86	27.36
✔	SAM	36.61	37.89	35.56	33.19	31.98
✔	SEM	38.36	38.52	36.74	**35.12**	**33.67**
✔	SNSep	**39.26**	**38.95**	**38.19**	34.94	31.56

4. Results and Discussions

Effectiveness of FTL on Audio Perception: Table 1 and Table 2 present the performance of LALMs on ASR and AT tasks, where Qwen3-8B is applied as the modality router.

For ASR, as shown in Table 1, FTL effectively reduces the WER of all evaluated LALMs, especially under low-SNR conditions where non-speech interference becomes dominant. For example, under the *SNR-Speech* of -10 dB, FTL reduces the WER of AF3 from 27.45% to 25.39%, and also improves FAC and Q3O by a similar margin.

For AT, Table 2 shows that FTL consistently improves mAP across different *SNR-Non-Speech* settings. The improvements are particularly notable when the target non-speech signals are contaminated by strong speech interference. For instance, under the *SNR-Non-Speech* of -10 dB, FTL improves the mAP of FAC from 16.34% to 20.75%.

These results suggest that our proposed FTL provides a general and effective mechanism for task-adaptive audio enhancement, benefiting both speech-centric and non-speech-centric audio perception tasks.

Effectiveness of FTL on Audio Reasoning: Table 3 presents the reasoning performance of FTL on MMAU-Pro-Ctrl under different modality routers and SNR conditions.

For speech reasoning, FTL generally improves QA-ACC under noisy conditions when accurate modality routing is available. For non-speech reasoning, FTL with ChatGPT5.2 and ground-truth routing also improves performance across most SNR settings. It should be noted that, Qwen3-8B achieves a CR of 0% because it consistently predicts "mixture", resulting in the original mixed audio being fed into the LALMs and thus providing no performance gain.

A surprising observation is that a better router does not always bring better reasoning performance. In some cases, the ground-truth router performs comparably to or even slightly worse than ChatGPT5.2 despite perfect routing accuracy. This suggests that, compared with low-level perception tasks, audio reasoning is more sensitive to signal distortion and contextual completeness, making the effectiveness of modality-aware enhancement less deterministic.

Impact of Audio Separator: In previous experiments, we use SNSep as the default separator. In Table 4, we further investigate the impact of different audio separators within the proposed FTL framework. In addition, the SDR performance of different separators is provided in Appendix C, where SNSep and SEM achieve comparable separation quality, significantly outperforming SAM.

As shown in Table 4, both SEM and SNSep consistently improve ASR and AT performance compared with the vanilla model, whereas SAM often leads to degraded ASR results. An interesting observation is that although SNSep achieves slightly better separation performance than SEM in terms of SDR, it does not yield better ASR improvements. This finding suggests that cleaner separation does not necessarily lead to better speech understanding for LALMs. Excessively aggressive separation may remove subtle acoustic cues or introduce distortions that negatively affect downstream perception. We provide a real-case analysis in Appendix D to further illustrate this phenomenon.

Performance on Real Mixtures: Since real mixtures lack ground-truth SNRs, we provide qualitative demonstrations of FTL on real-world mixtures on our project page[1]. Results show that FTL can also effectively improve audio understanding in real-world mixtures.

5. Conclusions

In this work, we propose FTL, an easy but efficient audio enhancement framework for noise-robust LALMs. Results show that FTL improves both audio perception and reasoning performance, especially under high-noise conditions, providing practical guidelines for deploying LALMs in real-world noisy scenarios. Despite its effectiveness, FTL applies a frozen LLM for modality routing; future work will study adaptive fusion and routing to improve robustness.

6. Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant (No. RS-2024-00337945); and the BK21 FOUR program through the NRF grant funded by the Ministry of Education of Korea government (MOE). This work was supported by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)&Gwangju Metropolitan City.

References

Chao, R., Cheng, W.-H., La Quatra, M., Siniscalchi, S. M., Yang, C.-H. H., Fu, S.-W., and Tsao, Y. An investigation of incorporating mamba for speech enhancement. In *Spoken Language Technology Workshop (SLT)*, pp. 302–308. IEEE, 2024.

Chen, G., Chai, S., Wang, G.-B., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *Interspeech*, 2021.

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. Vggsound: A large-scale audio-visual dataset. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., and Zhou, J. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Deshmukh, S., Elizalde, B., Singh, R., and Wang, H. Pengi: An audio language model for audio tasks. *Proc. Advances in Neural Information Processing Systems (NIPS)*, 36: 18090–18108, 2023.

Ding, D., Ju, Z., Leng, Y., Liu, S., Liu, T., Shang, Z., Shen, K., Song, W., Tan, X., Tang, H., et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Fonseca, E., Favory, X., Pons, J., Font, F., and Serra, X. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Ghosh, S., Kumar, S., Seth, A., Evuru, C. K. R., Tyagi, U., Sakshi, S., Nieto, O., Duraiswami, R., and Manocha, D. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6288–6313, 2024.

Goel, A., Ghosh, S., Kim, J., Kumar, S., Kong, Z., Lee, S.-g., Yang, C.-H. H., Duraiswami, R., Manocha, D., Valle, R., et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2025.

Gong, Y., Yu, J., and Glass, J. Vocalsound: A dataset for improving human vocal sounds recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 151–155. IEEE, 2022.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Hu, Y., Chen, C., Yang, C.-H. H., Li, R., Zhang, C., Chen, P.-Y., and Chng, E. Large language models are efficient learners of noise-robust speech recognition. *Proc. International Conference on Learning Representations (ICLR)*, 2024.

Ito, K. and Johnson, L. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

Jeong, I.-Y. and Park, J. Cochlscene: Acquisition of acoustic scene data using crowdsourcing. In *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 17–21. IEEE, 2022.

Kumar, S., Sedláček, Š., Lokegaonkar, V., López, F., Yu, W., Anand, N., Ryu, H., Chen, L., Plička, M., Hlaváček, M., et al. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *AAAI*, 2026.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, C.-A., Lin, T.-H., and Lee, H.-y. When silence matters: The impact of irrelevant audio on text reasoning in large audio-language models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17757–17761. IEEE, 2026.

Liu, X., Kong, Q., Zhao, Y., Liu, H., Yuan, Y., Liu, Y., Xia, R., Wang, Y., Plumbley, M. D., and Wang, W. Separate anything you describe. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 33:458–471, 2024.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2025.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. In *Proc. International conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Salamon, J., Jacoby, C., and Bello, J. P. A dataset and taxonomy for urban sound research. In *Proc. ACM International Conference on Multimedia (ACM MM)*, pp. 1041–1044, 2014.

Schmidt, M., Stier, D., Werner, S., and Minker, W. Exploration and assessment of proactive use cases for an in-car voice assistant. In *Konferenz elektronische sprachsignalverarbeitung*, pp. 148–155. TUDpress, Dresden, 2019.

Shi, B., Tjandra, A., Hoffman, J., Wang, H., Wu, Y.-C., Gao, L., Richter, J., Le, M., Vyas, A., Chen, S., et al. Sam audio: Segment anything in audio. *arXiv preprint arXiv:2512.18099*, 2025.

Team, T. F., Chen, Q., Cheng, L., Deng, C., Li, X., Liu, J., Tan, C.-H., Wang, W., Xu, J., Ye, J., et al. Fun-audio-chat technical report. *arXiv preprint arXiv:2512.20156*, 2025.

Torad, M. A., Bouallegue, B., and Ahmed, A. M. A voice controlled smart home automation system using artificial intelligent and internet of things. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 20(4):808–816, 2022.

Wang, B., Zou, X., Lin, G., Sun, S., Liu, Z., Zhang, W., Liu, Z., Aw, A., and Chen, N. Audiobench: A universal benchmark for audio large language models. In *Proc. the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4297–4316, 2025.

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proc. the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 993–1003, 2021.

Xie, Z., Lin, M., Liu, Z., Wu, P., Yan, S., and Miao, C. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*, 2025.

Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yin, H. and Choi, J.-W. Can large audio language models understand audio well? speech, scene and events understanding benchmark for lalms. *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

Yin, H., Wang, M., Bai, J., Shi, D., Gan, W.-S., and Chen, J. Sub-band and full-band interactive u-net with dprnn for demixing cross-talk stereo music. In *Proc. International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 21–22. IEEE, 2024.

Yin, S., Liu, C., Zhang, Z., Lin, Y., Wang, D., Tejedor, J., Zheng, T. F., and Li, Y. Noisy training for deep neural networks in speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):2, 2015.

Yu, W., Tang, C., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., and Zhang, C. Connecting speech encoder and large language model for asr. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12637–12641. IEEE, 2024.

Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y. J., and Ma, Y. Investigating the catastrophic forgetting in multimodal large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.

Zhang, G., Liang, H.-N., and Yue, Y. An investigation of the use of robots in public spaces. In *Proc. International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp. 850–855. IEEE, 2015.

Zhang, Y., Tian, J., Zhang, Y., Yan, S., Lin, L., Zhou, Z., Sun, L., and Su, S. See: Signal embedding energy for quantifying noise interference in large audio language models. *arXiv preprint arXiv:2601.07331*, 2026.

A. Prompts

A.1. Prompt for the LLM-based modality router in FTL

You are an expert in audio understanding and multimodal reasoning. Your task is to decide what audio input should be provided to a Large Audio Language Model (LALM) in order to best accomplish a user's instruction.

The audio has been separated into two tracks: speech: contains spoken voice content only; non-speech: contains non-speech acoustic events only. Mixture refers to the original unseparated audio.

You should select the input that maximizes task-relevant information, based on the user's instruction.

Guidelines:

1. You should ONLY choose ‘speech’ when speech information alone is clearly sufficient to solve the task, AND non-speech provides no meaningful additional information.

2. You should ONLY choose ‘non-speech’ when non-speech audio alone is clearly sufficient to solve the task, AND speech provides no meaningful additional information.

3. In ALL other cases, including uncertainty, partial usefulness of both modalities, or when you cannot strictly rule out one modality, you MUST choose ‘mixture’.

Additional Domain Rules: - Speech is required for linguistic content, speaker intent, emotion, or dialogue understanding. - Non-speech includes environmental sounds and vocal non-linguistic sounds (e.g., laughter, sneeze, cough).

Respond with only one word: speech, non-speech, or mixture. Do not provide explanations.

User Instruction: [**the user's instruction**].

A.2. User's Instructions

A.2.1. AUTOMATIC SPEECH RECOGNITION (ASR) TASK

Transcribe the speech into text, without any further explanation.

A.2.2. AUDIO TAGGING (AT) TASK

You are an expert in sound events classification. I will give you an audio recording. Please carefully analyze the sound events in this audio. Ignore speech and focus only on non-speech sound events. Output only one line, no explanations. List events detected in the audio, separated by a semicolon and a space. If no event is detected, output: None.

B. Ablations of Hyper-parameters

In this section, we analyze the impact of the fusion coefficients α_{sp} and α_{ns} on the performance of FTL, and explain how the default values ($\alpha_{sp} = 0.5$ and $\alpha_{ns} = 0.9$) are determined. Table A1 reports the ASR and AT results of three LALMs (AF3, FAC, and Q3O) under different α_{sp} and α_{ns} values on SSEU-Bench. Figure A1 further visualizes the impact of α_{sp} on ASR at three representative SNR-Speech conditions.

Impact of α_{sp} on ASR: As shown in Table A1 and Figure A1, the WER of all three LALMs exhibits a clear U-shaped trend with respect to α_{sp}, especially under low-SNR conditions. When $\alpha_{sp} = 1.0$, i.e., the LALM directly consumes the separated speech signal, the WER becomes substantially worse than the vanilla baseline. For example, at SNR-Speech $= -10$ dB, $\alpha_{sp} = 1.0$ degrades the WER of AF3 from 27.45% to 37.50%, and FAC from 31.67% to 44.41%. This degradation indicates that, although fully replacing the input with the separated speech effectively removes non-speech interference, the separator inevitably introduces artifacts and spectral distortions that mislead the acoustic perception of LALMs. On the other hand, when $\alpha_{sp} = 0.1$, the enhanced signal is dominated by the raw mixture, providing only marginal suppression of non-speech interference and thus limited improvement over the baseline. The setting of $\alpha_{sp} = 0.5$ achieves the best overall balance: it removes a substantial portion of non-speech interference while retaining sufficient natural acoustic context from the raw audio to preserve signal fidelity. As a result, $\alpha_{sp} = 0.5$ consistently yields the lowest WER across all three LALMs and most SNR levels (e.g., AF3 from 27.45% to 25.39% and Q3O from 20.42% to 18.61% at SNR-Speech $= -10$ dB), and is therefore selected as the default value.

Table A1. Performance of different LALMs on SSEU-Bench. "SNR-Speech" denotes the speech-to-non-speech ratio; "SNR-Non-Speech" refers to non-speech to speech ratio.

LALM	FTL	α_{sp}	ASR Performance, Metric: WER (%)						α_{ns}	AT Performance, Metric: mAP (%)					
			SNR-Speech (dB)							SNR-Non-Speech (dB)					
			$+\infty$	10	5	0	-5	-10		$+\infty$	10	5	0	-5	-10
AF3	✗	-	2.18	2.71	3.27	4.73	10.45	27.45	-	38.80	36.18	34.56	31.00	28.86	27.36
	✔	1.0	2.21	3.13	3.93	6.32	15.93	37.50	1.0	39.22	38.55	37.43	36.44	34.86	**31.94**
	✔	0.9	**2.15**	2.89	3.49	5.45	12.29	31.15	0.9	**39.28**	**39.26**	**38.95**	**38.19**	**34.94**	31.56
	✔	0.5	2.17	**2.66**	**3.20**	**4.61**	**9.83**	**25.39**	0.5	39.16	36.52	35.34	32.92	31.24	29.29
	✔	0.1	2.16	2.70	3.43	4.63	9.93	26.73	0.1	39.19	36.06	33.01	31.51	29.34	27.70
FAC	✗	-	2.61	3.38	3.99	5.75	12.54	31.67	-	36.34	21.27	18.33	17.54	16.98	16.34
	✔	1.0	2.82	3.66	4.94	8.47	20.38	44.41	1.0	36.34	**33.22**	**31.73**	**32.32**	**31.77**	**29.30**
	✔	0.9	**2.58**	3.37	4.26	6.69	15.20	35.63	0.9	**36.64**	31.97	30.32	27.88	24.89	20.75
	✔	0.5	2.61	**3.24**	**3.86**	**5.44**	**11.54**	**28.41**	0.5	36.16	26.62	21.10	18.43	17.74	17.39
	✔	0.1	**2.58**	3.32	3.91	5.71	12.00	30.78	0.1	36.60	21.80	18.58	17.78	17.42	16.31
Q3O	✗	-	2.16	2.31	2.56	3.64	7.04	20.42	-	44.43	39.75	38.12	34.84	33.20	31.33
	✔	1.0	**2.05**	**2.18**	2.66	3.99	9.33	29.12	1.0	**44.66**	**43.87**	**43.46**	**42.01**	**39.94**	**37.30**
	✔	0.9	2.14	2.45	2.66	3.55	7.86	23.75	0.9	44.27	43.48	42.25	40.32	39.20	37.27
	✔	0.5	2.23	2.31	**2.49**	**3.38**	**5.97**	**18.61**	0.5	44.38	40.49	38.65	37.58	34.76	32.97
	✔	0.1	2.20	2.38	2.58	3.55	6.82	19.94	0.1	44.55	40.31	37.83	35.88	33.13	30.98

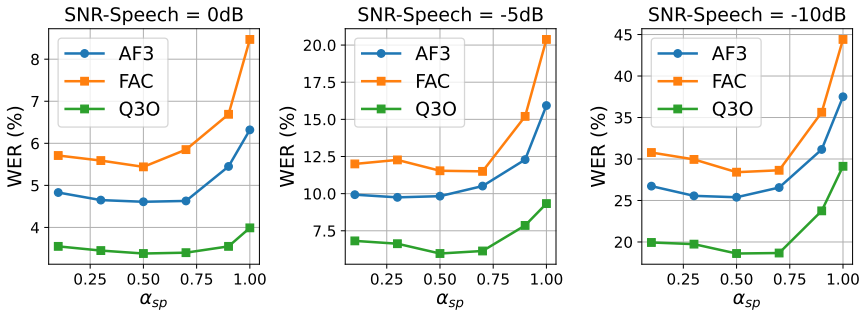

Figure A1. Impact of α_{sp} on ASR task of SSEU-Bench.

Impact of α_{ns} on AT: Unlike ASR, the AT performance generally improves as α_{ns} increases. As shown in Table A1, when $\alpha_{ns} = 1.0$, the LALM receives the purely separated non-speech signal and achieves strong AT performance across all LALMs and SNR conditions. For instance, at SNR-Non-Speech $= -10$ dB, $\alpha_{ns} = 1.0$ improves the mAP of AF3 from 27.36% to 31.94%, and Q3O from 31.33% to 37.30%. Setting $\alpha_{ns} = 0.9$ produces results highly comparable to $\alpha_{ns} = 1.0$, and even surpasses it in several cases (e.g., AF3 achieves 39.26% mAP at SNR-Non-Speech $= 10$ dB with $\alpha_{ns} = 0.9$, higher than 38.55% with $\alpha_{ns} = 1.0$). In contrast, smaller α_{ns} values (0.5 and 0.1) yield notably weaker AT performance, since the enhanced signal still contains a large portion of speech interference that distracts the LALMs from non-speech.

Although $\alpha_{ns} = 1.0$ slightly outperforms $\alpha_{ns} = 0.9$ on the AT task in some settings, we adopt $\alpha_{ns} = 0.9$ as the default value for the following robustness consideration. When $\alpha_{ns} = 1.0$, the enhanced signal consists exclusively of the separated non-speech component, meaning that the speech track is completely discarded. In practice, the modality router is not always perfect: if the router misjudges a speech-related instruction as "non-speech", such an aggressive setting would eliminate all spoken content from the input, leading to a catastrophic failure on the downstream task. Retaining 10% of the raw audio ($\alpha_{ns} = 0.9$) ensures that even under occasional routing errors, the speech information is partially preserved, preventing complete loss of task-relevant content. This conservative design trades a small amount of AT performance for substantially better robustness against router errors, which we believe is a favorable trade-off for real-world deployment.

Summary: Based on the above analysis, we adopt $\alpha_{sp} = 0.5$ and $\alpha_{ns} = 0.9$ as the default hyper-parameters in FTL. These values are chosen not only for their strong empirical performance, but also to jointly balance enhancement strength, signal fidelity, and robustness against potential router misjudgments.

C. Performance of Audio Separators

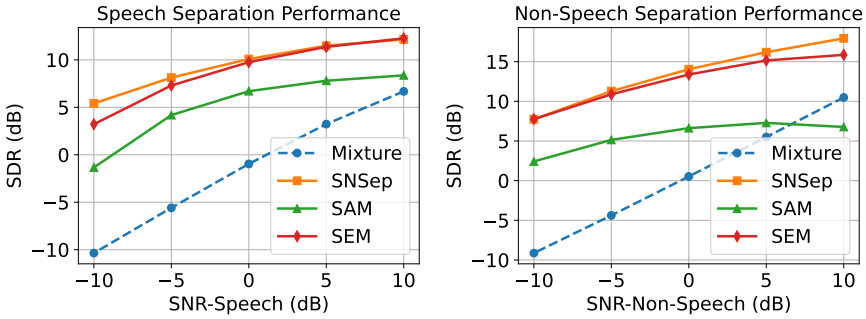

Figure A2. Speech and non-speech separation performance of different separators on SSEU-Bench.

In this section, we evaluate the separation quality of three audio separators considered in FTL, i.e., SNSep, SAM, and SEM. Figure A2 reports the Signal-to-Distortion Ratio (SDR) of each separator on the speech and non-speech tracks of SSEU-Bench, under varying SNR conditions. As a reference, we also include the SDR of the raw mixture, which reflects the difficulty of the input.

Overall Comparison: As shown in Figure A2, all three separators substantially outperform the raw mixture baseline on both speech and non-speech tracks, confirming that explicit separation provides meaningful signal-level improvement for both target modalities. Among the three separators, SNSep and SEM consistently achieve the highest SDR across all SNR levels, while SAM lags behind by a noticeable margin. For instance, at SNR-Speech $= -10$ dB, SNSep and SEM yield speech SDRs above 3 dB, whereas SAM only reaches around -1 dB. A similar trend is observed on the non-speech track, where SNSep and SEM clearly outperform SAM by roughly 5 dB at low-SNR conditions.

Comparison between SNSep and SEM: SNSep and SEM achieve highly comparable separation quality in terms of SDR, with SNSep slightly better in most SNR conditions, particularly on the non-speech track at low SNRs. This is consistent with our design motivation: SEM is a speech enhancement model that estimates non-speech by subtracting the enhanced speech from the mixture, which inevitably propagates speech-estimation errors into the non-speech branch. In contrast, SNSep adopts a dedicated dual-decoder architecture that independently reconstructs the speech and non-speech tracks, leading to a more balanced separation across both modalities.

Discussion: The SDR results above explain the trends observed in Table 4 of the main paper. Specifically, both SNSep and SEM achieve consistent improvements on ASR and AT when used in FTL, while SAM occasionally degrades the ASR performance due to its inferior separation quality. Notably, although SNSep yields slightly higher SDR than SEM, it does not always translate into better downstream ASR performance. This observation indicates that signal-level metrics such as SDR are not perfectly aligned with the perceptual preferences of LALMs, and that excessive separation may remove subtle acoustic cues useful for downstream understanding. We adopt SNSep as the default separator in FTL because of its overall balanced separation quality across both speech and non-speech tracks, while leaving the joint optimization of separation and downstream LALM perception as an interesting direction for future work.

D. Real-Case Analysis

To better understand why a separator with higher SDR does not always lead to better downstream ASR performance, we present a representative case study in Figure A3. The example consists of a clean speech utterance and a non-speech sound mixed at SNR-Speech $= 5$ dB. We feed five different versions of this sample into Audio Flamingo 3 (AF3) and report both the SDR (against the clean speech reference) and the corresponding WER.

SDR is not Always Aligned with ASR Performance: As shown in Figure A3, the raw mixture yields a low SDR of 3.52 dB and causes AF3 to produce a WER of 10.71%. After separation, both SEM and SNSep substantially increase the SDR (7.99 dB and 8.48 dB, respectively), confirming that they effectively remove the non-speech component at the signal level. However, the downstream ASR results tell a different story: SEM enables AF3 to perfectly transcribe the utterance (WER $= 0.00\%$), whereas SNSep, despite having the highest SDR, still leads to a non-trivial WER $= 3.57\%$.

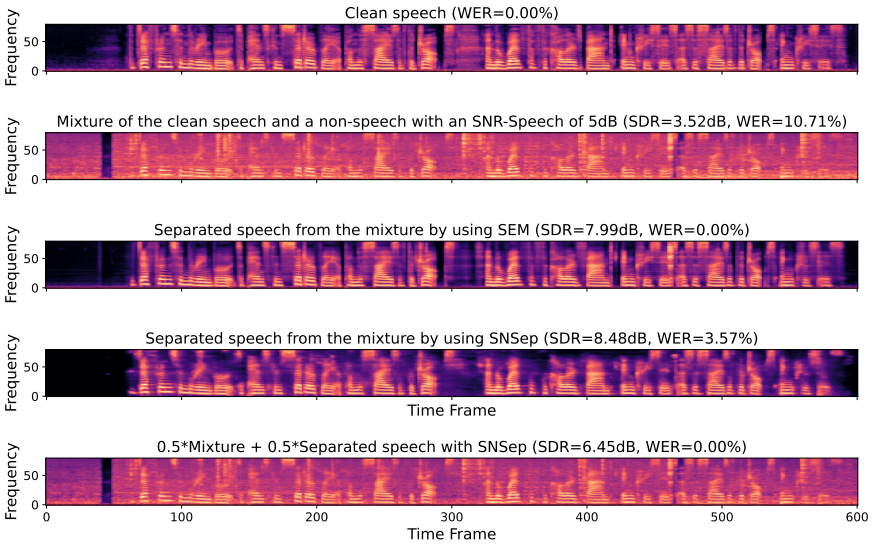

Figure A3. ASR demonstration (mel spectrogram): Each sample is fed into Audio Flamingo 3 to perform ASR.

This counter-intuitive result aligns with our observation in Section 4 of the main paper, namely that signal-level separation quality does not directly translate into perceptual benefits for LALMs.

Why Does SNSep Hurt ASR? A closer inspection of the mel spectrograms reveals the underlying cause. Compared with the clean speech reference and the SEM-separated output, the SNSep-separated speech exhibits a clearly attenuated low-frequency band at the bottom of the mel spectrogram (highlighted by the missing horizontal energy strip in Figure A3). This region typically carries vocal fundamental frequencies and low-order harmonics that are crucial for natural speech perception. Although removing such energy bands helps suppress residual non-speech interference (and thus yields a slightly higher SDR), it also makes the resulting waveform acoustically unnatural and out-of-distribution from the perspective of the pre-trained LALM. As a result, AF3 misrecognizes some words despite the higher SDR. This phenomenon illustrates that overly aggressive separation can sacrifice acoustic naturalness, which is more important than raw signal fidelity for downstream LALM perception.

How MAFB Mitigates the Issue: The bottom panel of Figure A3 shows the output of our MAFB with $\alpha_{sp} = 0.5$, i.e., the linear combination of the raw mixture and the SNSep-separated speech. Although this fusion lowers the SDR to 6.45 dB (since the mixture re-introduces a small amount of non-speech interference), it restores the missing low-frequency band and recovers the natural acoustic structure of speech. Crucially, AF3 now achieves a perfect WER $= 0.00\%$ on this fused signal. This case study provides direct empirical support for the design of MAFB: by mixing the separated signal with the raw audio, MAFB compensates for the artifacts and missing acoustic cues introduced by aggressive separation, achieving a better trade-off between noise suppression and signal naturalness than relying on the separator output alone.