EXAM^2: Extending Audio Understanding in Multilingual and Multimodal Analysis

Jiawen Wang[1] **Xiaoxue Gao**[2†] **Zi Haur Pang**[3] **Nancy F. Chen**[4]

[1]University of Munich, Germany
[2]School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, China
[3]Kyoto University, Japan [4]A*STAR, Singapore
jiawen.wang@campus.lmu.de

Abstract

Recent large audio language models (LALMs) have achieved impressive progress in audio understanding. However, existing evaluations remain largely constrained to English and narrow audio domains. Prior benchmarks typically focus on a single audio modality, i.e., speech, sound, or music, limiting the systematic investigation into how these models generalize across diverse visual scenarios. In this paper, we introduce EXAM^2, a benchmark for multilingual and multimodal audio understanding spanning six languages and multiple modalities, including speech, sound, music, mixed-audio settings, and visual images. By incorporating visual information alongside heterogeneous audio inputs, EXAM^2 enables more realistic evaluation of scene-aware audio reasoning and cross-modal comprehension. EXAM^2 comprises $5,667$ multiple-choice questions, $22,614$ image instances, and $135,684$ multilingual translations. We evaluate state-of-the-art open-source and proprietary LALMs as well as multimodal LLMs, revealing substantial performance gaps in multilingual and cross-modal understanding. Furthermore, we propose Gemma3n-EXAM^2, a lightweight fusion-model fine-tuned on EXAM^2-train, achieves up to 15.8% improvement in multilingual settings and 16.5% gains in multimodal evaluation over a strong baseline. Empirical results establish EXAM^2 as a challenging benchmark and pioneer future multilingual and multimodal audio intelligence research.[1]

1 Introduction

Recent advancements of large audio language models (LALMs) (Deshmukh et al., 2023; Chu et al., 2024; Kong et al., 2024; Ghosh et al., 2025, 2026) has significantly expanded machine capabilities for understanding, reasoning over, and interacting with auditory signals (Wang et al., 2025b; Pang et al., 2026). To assess these capabilities, multiple choice question (MCQ) tasks (He et al., 2025) are being utilized across distinct audio modalities, including speech (Zhao et al., 2024; Huang et al., 2024), sound (Lipping et al., 2022; Ghosh et al., 2024), and music (Melechovsky et al., 2024; Weck et al., 2024), to evaluate how models interpret and reason about audio content under various contexts and scenarios. Consequently, a wide array of benchmarks have emerged, including AudioBench (Wang et al., 2025a), MMAU (Sakshi et al., 2025), and MMAR (Ma et al., 2025). Yet, these benchmarks remain primarily audio-only and rarely investigate interactions between heterogeneous audio sources and visual grounding (Yue et al., 2024, 2025; Wang et al., 2025c; Chen et al., 2025).

Beyond cross-modal limitations, restricted language coverage remains a critical bottleneck in current audio understanding evaluations. However, the semantic interpretation of audio content is frequently intertwined with linguistic and cultural contexts; for instance, identical lyrical slang can convey different meanings in diverse languages (Werner, 2023). Yet, existing audio-understanding benchmarks largely rely on English-centric question-answer settings (Drossos et al., 2020; Yang et al., 2024; Sakshi et al., 2025), overlooking models' capabilities to navigate these shifting contexts. This discrepancy underscores the necessity for multilingual question-answer benchmarks to facilitate a broader assessment of audio comprehension across diverse language settings.

To address these gaps, we introduce EXAM^2, a benchmark for multilingual and multimodal audio understanding under a unified MCQ evaluation framework (cf. Figure 1). EXAM^2 extends conventional audio-only benchmarks by jointly combining speech, sound, music, mixed-audio scenarios,

† Corresponding author.

[1]All code, data, and models are available on https://github.com/werywjw/EXAM-2.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2527–2543
November 6-10, 2026 ©2026 Association for Computational Linguistics



Figure 1: Overview of a multimodal and multilingual example from our EXAM^2 Benchmark.

and visual representations across six languages: English, German, Spanish, Japanese, Malay, and Chinese. Beyond measuring recognition accuracy, our benchmark probes multilingual reasoning, semantic grounding, and cross-modal comprehension in contemporary LALMs, multimodal large language models (MLLMs), and cascaded systems via large language models (LLMs). Finally, our proposed lightweight Gemma3n-EXAM^2 model achieves up to $+15.80\%$ average multilingual improvement ($+22.32\%$ on the speech category) and 16.50% gains in multimodal settings compared to the baseline, demonstrating the model's effectiveness for advancing multilingual and multimodal audio MCQ performance.

Our contributions mainly include: (1) We propose the EXAM^2 benchmark, to our best knowledge, the first to address multilingual and multimodal audio understanding with $5,667$ MCQs, $22,614$ image instances, and $135,684$ multilingual translated instances; (2) We introduce OmniLoRA, a straightforward omni-language uniform tuning based on LoRA, extensive experiments show the effectiveness of our method; (3) We present our multilingual Gemma3n-EXAM^2, which is a lightweight fusion approach with significant improvements across languages to foster future research in multilingual and multimodal domains; (4) We conduct extensive experiments using state-of-the-art LLMs and MLLMs on the EXAM^2 benchmark, offering a more holistic evaluation of audio intelligence and reasoning capabilities.

2 Related Work

Large Audio Language Models. Early works such as AudioCLIP (Guzhov et al., 2022) and CLAP (Elizalde et al., 2023) focused on learning shared audio-text representations for retrieval and captioning tasks. Building on these foundations, LALMs such as Qwen2-Audio (Chu et al., 2024), Audio Flamingo (Kong et al., 2024; Ghosh et al., 2025, 2026), and SALMONN (Tang et al., 2024) integrate audio encoders with LLMs to support instruction following and complex audio understanding. In parallel, general-purpose omni-models such as Qwen-omni (Xu et al., 2025; Team, 2026), Ming-omni (AI et al., 2025) and Baichuan-omni (Li et al., 2025) have shown strong cross-modal generalization despite not being specifically designed for audio tasks. While prior benchmarks mainly focus on evaluating LALMs, our work provides a comprehensive benchmark across LALMs, MLLMs, and cascaded LLMs that first transcribe audio before performing reasoning with (large) language or reasoning models.

Multilingual and Multimodal Audio Understanding Benchmarks. Previous benchmarks

Benchmark	Text	Audio	Vision	Language					
	Tr / Ca	Sp / So / Mu / Mix	Image	EN	DE	ES	JA	MS	ZH
Clotho	✗ / ✓	✓ / ✓ / ✓ / ✗	✗	✓	✗	✗	✗	✗	✗
Clotho-AQA	✗ / ✗	✗ / ✓ / ✗ / ✗	✗	✓	✗	✗	✗	✗	✗
LibriSQA	✗ / ✗	✓ / ✗ / ✗ / ✗	✗	✓	✗	✗	✗	✗	✗
Dynamic-SUPERB	✗ / ✗	✓ / ✗ / ✗ / ✗	✗	✓	✗	✗	✗	✗	✗
MusicBench	✗ / ✗	✗ / ✗ / ✓ / ✗	✗	✓	✗	✗	✗	✗	✗
MuChoMusic	✗ / ✗	✗ / ✗ / ✓ / ✗	✗	✓	✗	✗	✗	✗	✗
AIR-Bench	✗ / ✗	✓ / ✓ / ✓ / ✓	✗	✓	✗	✗	✗	✗	✗
AudioSetCaps	✗ / ✓	✓ / ✓ / ✓ / ✗	✗	✓	✗	✗	✗	✗	✗
AudioMCQ	✗ / ✗	✓ / ✓ / ✓ / ✗	✗	✓	✗	✗	✗	✗	✗
MMAR	✗ / ✗	✓ / ✓ / ✓ / ✓	✗	✓	✓	✗	✓	✗	✓
MMAU	✗ / ✗	✓ / ✓ / ✓ / ✗	✗	✓	✗	✗	✗	✗	✗
MMAU-Pro	✗ / ✗	✓ / ✓ / ✓ / ✗	✗	✓	✗	✗	✗	✗	✗
MMMU	✗ / ✗	✗ / ✗ / ✗ / ✗	✓	✓	✗	✗	✗	✗	✗
MMAU-Pro	✗ / ✗	✗ / ✗ / ✗ / ✗	✓	✓	✗	✗	✗	✗	✗
EXAM^2 (Ours)	✓ / ✓	✓ / ✓ / ✓ / ✓	✓	✓	✓	✓	✓	✓	✓

Table 1: The list of existing multimodal and multilingual datasets show the novelty of our EXAM^2 benchmark. Tr, Ca denotes Transcript and Caption, respectively. Sp / So / Mu / Mix stands for Speech, Sound, Music, and their Mix.

Benchmark	Statistics	Number
EXAM^2-test	Question	998
	Audio Domains	3
	Domain Categories	Speech / Sound / Music
	Category Distribution	333:333:332 (1:1:1)
	Visual Choices	3,938
	Languages	EN / DE / ES / JA / MS / ZH
	Multilingual Instances	23,628
EXAM^2-train	Question	4669
	Audio Domains	4
	Domain Categories	Speech / Sound / Music / Mix
	Category Distribution	396:3737:243:293
	Visual Choices	18,676
	Languages	EN / DE / ES / JA / MS / ZH
	Multilingual Instances	112,056
EXAM^2	Question	5667
	Audio Domains	4
	Domain Categories	Speech / Sound / Music / Mix
	Category Distribution	729:4070:575:293
	Visual Choices	22,614
	Languages	EN / DE / ES / JA / MS / ZH
	Multilingual Instances	135,684
	Average Question Length	10.7 words
	Average Option Length	3.2 words
	Average Audio Duration	12.9 sec

Table 2: Dataset statistics of our EXAM^2 benchmark.

have advanced the evaluation of LALMs across speech, sound, and music understanding. Domain specific benchmarks such as LibriSQA (Zhao et al., 2024) and Dynamic-SUPERB (Huang et al., 2024) for speech-related tasks, Clotho-AQA (Lipping et al., 2022), CompA (Ghosh et al., 2024) for environmental sound reasoning (Iyer et al., 2026), and MusicBench (Melechovsky et al., 2024) and MuChoMusic (Weck et al., 2024) for music understanding. Broader benchmarks such as AudioBench (Wang et al., 2025a), AIR-Bench (Yang et al., 2024), MMAU (Sakshi et al., 2025), MMAU-Pro (Kumar et al., 2026), and MMAR (Ma et al., 2025) further combine multiple audio domains and evaluate advanced auditory reasoning capabilities. However, existing benchmarks remain largely audio-only, with limited support for multimodal grounding or visual representations in audio understanding tasks. In addition, multilingual evaluation is still underexplored. Prior multilingual benchmarks such as BUFFET (Asai et al., 2024) and MEGA (Ahuja et al., 2023) focus on text-based LLMs rather than the MLLMs, while the audio benchmark MMAR provides very limited multilingual audio and entirely lacks multilingual question-and-answer pairs.

3 EXAM^2 Benchmark

3.1 Overview

Our EXAM^2 introduces multilingual and multimodal audio understanding with visual representations, supporting six languages (DE, ES, JA, MS, ZH, and EN) to enable richer cross-modal and cross-lingual reasoning (cf. Table 1). Dataset statistics for our proposed EXAM^2 benchmark are given in Table 2. The benchmark consists of two main subsets: EXAM^2-train and EXAM^2-test, with a total of $5,667$ audio instances, $135,684$ multilingual language instances, and $22,614$ visual answer choices. EXAM^2-train, recreated and filtered from MMAR and Clotho, includes $4,669$ questions covering four audio domains (speech, sound, music, and mix) with $18,676$ visual choices. EXAM^2-test, restructured and annotated from MMAU-test-mini, contains 998 questions across three audio domains (speech, sound, and music) with a balanced category distribution and a total of $3,938$ visual choices.

3.2 Dataset Construction

Figure 2 illustrates the construction pipeline of our EXAM^2 benchmark with seven key steps.

Source Selection. We begin by collecting diverse audio corpora spanning speech, music, and environmental sounds, prioritizing real-world recordings over synthetic data to ensure ecological validity. Specifically, we curate four representative datasets: MMAU-test-mini (Sakshi et al., 2025), MMAR (Ma et al., 2025), Clotho (Drossos et al., 2020), and AudioMCQ-Clotho (He et al., 2025) under `Creative Commons` license to ensure a strong foundation for task development.

Quality Control and Filtering. To ensure benchmark quality, all collected audio samples, textual questions, choices, and answers undergo manual inspection by the authors. For MMAU-test-mini,

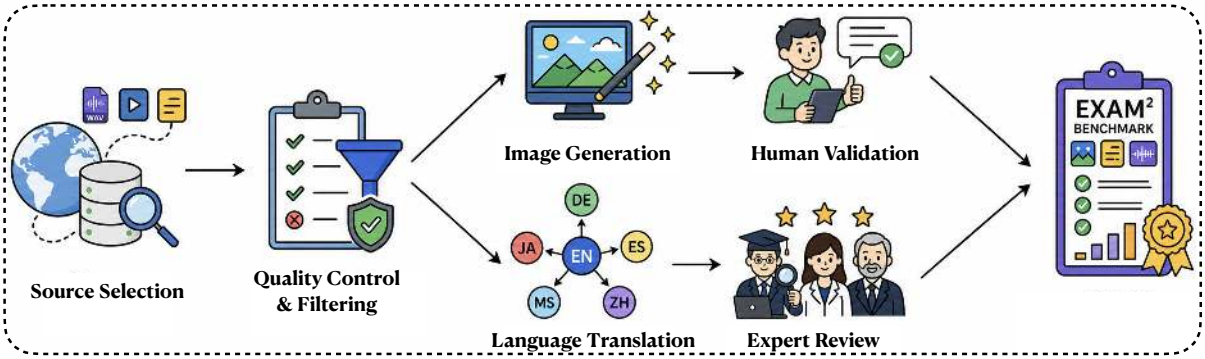


Figure 2: EXAM^2 Benchmark Construction Pipeline.

we remove two music-category samples that fail to meet the quality standards and eliminate duplicate answer choices within the same question, resulting in a final test set of 998 instances with 3,938 answer choices. For MMAR, we retrieve audio samples from the provided URLs and discard 129 out of 1,000 instances due to inaccessible or broken links, yielding 871 valid audio instances with 3,484 answer choices. For the Clotho benchmark, we derive multiple-choice questions from AudioMCQ-Clotho, obtaining 3,798 audios with 15,192 choices.

Image Generation. To enable multimodal evaluation, we generate visual counterparts for all answer choices in EXAM^2. For the smaller subsets, EXAM^2-MMAU and EXAM^2-MMAR, we employ Stable Diffusion (Rombach et al., 2022) for image generation under `CreativeML Open RAIL-M` license. For EXAM^2-Clotho, visual answer choices are generated using GPT-image-2.[2]

Human Validation. All textual questions and generated visual candidates are manually reviewed by the experts to ensure semantic relevance, clarity, and overall quality. When generated images are judged to be misleading, ambiguous, or insufficiently aligned with the corresponding answer choice, we manually modified 25.05% of the generated images to real-world screenshots to ensure the visual quality. Representative examples across categories are provided in Appendix A.2.2.

Language Translation. To support multilingual evaluation, we translate the original English questions and answer choices into five target languages: German (DE), Spanish (ES), Japanese (JA), Malay (MS), and Chinese (ZH). We employ GPT-5-mini[3] for translation, using carefully designed prompts (See Appendix A.3) to preserve semantic fidelity and task consistency across languages. Domain-specific terminology related to audio understanding is further reviewed and manually refined where necessary to ensure accurate translation and conceptual equivalence.

Expert Review. Translated questions and answer choices are reviewed by either native speakers of the target languages or experienced language users with more than five years of advanced proficiency. The review process focuses on semantic correctness, linguistic fluency, and preservation of task intent across languages, ensuring high-quality multilingual question-answer pairs across all benchmark languages. The manual edit rates of translation are 3.01% (DE), 2.30% (ES), 2.81% (JA), 1.10% (MS), and 5.31% (ZH). An inter-annotator agreement of Cohen's kappa achieves 0.9830, indicating almost perfect agreement.

EXAM^2 Finalization. The final EXAM^2 benchmark comprises 5,667 audio instances, 135,684 multilingual language instances, and 22,614 visual answer choices. The benchmark is designed to provide balanced coverage of speech, sound, and music understanding across six languages (EN, DE, ES, JA, MS, and ZH), enabling comprehensive evaluation of multimodal and multilingual audio understanding.

4 Methodology

4.1 Task Formulation

We address multimodal MCQ over audio and visual inputs. Each instance is a tuple $(\mathbf{a}, \mathcal{V}, q, \mathcal{C}, y^*)$,

[2] https://openai.com/index/introducing-chatgpt-images-2-0/

[3] https://openai.com/index/introducing-gpt-5/

where $\mathbf{a} \in \mathbb{R}^T$ is a raw audio waveform of T samples, $\mathcal{V} = \{v_1, \ldots, v_K\}$ is a set of K candidate images (one per choice), q is a natural-language question, $\mathcal{C} = \{c_1, \ldots, c_K\}$ is the set of candidate answers, and $y^* \in \mathcal{C}$ is the ground-truth answer. The model must identify y^* by jointly reasoning over both modalities.

4.2 Multimodal Input Representation

We build on Gemma3n-E4B[4], a MLLM that encodes audio, images, and text within a unified transformer.

Audio. The waveform $\mathbf{a}$ is resampled to 16 kHz and converted to a spectrogram $\mathbf{S} \in \mathbb{R}^{F \times L}$ (F frequency bins, L frames). A convolutional audio encoder projects $\mathbf{S}$ into a sequence of d-dimensional hidden states $\mathbf{H}^{(a)} \in \mathbb{R}^{L' \times d}$, which are prepended to the token sequence as soft audio tokens.

Images. For each choice c_k, a corresponding image $v_k \in \mathbb{R}^{H \times W \times 3}$ is encoded into patch embeddings $\mathbf{H}^{(v_k)} \in \mathbb{R}^{N_p \times d}$, where N_p is the number of patches. The image embeddings for all K choices are concatenated and interleaved with textual tokens.

Multimodal Fusion. The combined input sequence fed to the language-model backbone is

$$\mathbf{X} = [\,\mathbf{H}^{(v_1)}, \ldots, \mathbf{H}^{(v_K)},\ \mathbf{H}^{(a)},\ \mathbf{e}_q,\ \mathbf{e}_{\mathcal{C}}\,], \quad (1)$$

where $\mathbf{e}_q$ and $\mathbf{e}_{\mathcal{C}}$ are the token embeddings of the question and choice text, respectively. No modality-specific fusion layer is added; cross-modal alignment is learned entirely through self-attention.

4.3 Parameter-Efficient Fine-Tuning via LoRA

Full fine-tuning of all model parameters θ is computationally prohibitive. Following Hu et al. (2022), we freeze θ and inject trainable low-rank update matrices into the attention and feed-forward layers. We fine-tune the model with a cross-entropy loss restricted to the answer token. Given the multimodal input $\mathbf{X}$ (Eq. 1), the model is trained to predict the single-digit index $y^* \in \{1, \ldots, K\}$ identifying the correct choice:

$$\mathcal{L}(\theta_{\text{LoRA}}) = -\log P_\theta(y^* \mid \mathbf{X})\,. \quad (2)$$

After training, the LoRA adapters are merged back into the base weights and saved as a single self-contained model for inference. To support multilingual evaluation, we propose omni-language uniform tuning (OmniLoRA), which applies LoRA to the backbone model by uniformly sampling one of six languages $\ell \in \{\texttt{EN}, \texttt{DE}, \texttt{ES}, \texttt{JA}, \texttt{MS}, \texttt{ZH}\}$ for each training instance per epoch:

$$\mathcal{L}_{\text{Omni}}(\theta_{\text{LoRA}}) = -\,\mathbb{E}_{\ell \sim \mathcal{L}}\Big[\log P_\theta\Big(y^* \mid \mathbf{X}^{(\ell)}\Big)\Big]\,, \quad (3)$$

where $\mathbf{X}^{(\ell)}$ denotes the fused multimodal sequence (Eq. 1) with text rendered in language ℓ. This design keeps pre-trained multimodal representations intact while efficiently adapting the model to the target task across six languages.

5 Evaluation

5.1 Experimental Setup

We evaluate a diverse set of closed- and open-source models on the proposed EXAM^2 benchmark. Our results can be easily reproduced using the provided codebase and evaluation scripts. Closed-source models and Phi-4-multimodal-instruct are conducted using the Azure OpenAI API, while the remaining open-source models are evaluated using Hugging Face library. The supervised fine-tuning on our EXAM^2-train allocates four NVIDIA A40 GPUs. For all models and languages, we use the same set of prompts, detailed in Appendix A.3, across audio domains (speech, sound, and music) to ensure a fair comparison. Note that GPT-4o-transcribe[5] is deployed for audio transcription tasks.

5.2 Baseline

Closed- and open-source models are evaluated on the EXAM^2-test, in zero-shot settings. For closed-source models, we evaluate GPT-4o-audio[6] and GPT-4o-mini-audio[7]. GPT-4o[8] is leveraged as a MLLM with image input, while GPT-5-mini[9] and GPT-5.2[10] are evaluated as text-only mod-

[4] https://huggingface.co/google/gemma-3n-E4B-it

[5] https://developers.openai.com/api/docs/models/gpt-4o-transcribe

[6] https://developers.openai.com/api/docs/models/gpt-4o-audio-preview

[7] https://developers.openai.com/api/docs/models/gpt-4o-mini-audio-preview

[8] https://developers.openai.com/api/docs/models/gpt-4o

[9] https://openai.com/gpt-5/

[10] https://openai.com/index/introducing-gpt-5-2/

Model	Input Modality		Average (%)			
	Image	Audio	Speech	Sound	Music	Avg.
Closed-source						
GPT-4o-mini-audio	✘	✔	73.67	61.01	57.13	63.94
GPT-4o-audio	✘	✔	77.93	73.17	70.38	73.83
GPT-4o†	✔	✓	66.67	69.17	57.13	64.32
GPT-5-mini†	✘	✓	71.82	51.25	54.29	59.12
GPT-5-2†	✘	✓	71.47	50.25	50.10	57.27
Open-source						
Phi-4-multimodal-instruct	✘	✔	<u>59.91</u>	54.47	55.02	56.47
Phi-4-multimodal-instruct†	✘	✓	47.05	60.81	56.21	54.69
Phi-4-multimodal-instruct†	✔	✓	51.30	59.91	44.08	51.76
Qwen-2-audio	✘	✔	34.29	34.33	40.36	36.33
Qwen-2.5-omni†	✘	✓	53.60	64.17	<u>56.52</u>	58.10
Qwen-2.5-omni†	✔	✓	53.25	61.11	55.37	56.58
Qwen-2.5-omni	✘	✔	39.64	49.90	32.58	40.71
Qwen-2.5-omni	✔	✔	52.20	61.61	54.97	56.26
Gemma3n†	✘	✓	58.41	**67.12**	54.22	<u>59.91</u>
Gemma3n†	✔	✓	57.16	<u>66.37</u>	53.46	59.00
Gemma3n	✔	✘	42.14	48.70	46.64	45.83
Gemma3n	✘	✔	42.84	43.04	48.34	44.74
Gemma3n	✔	✔	41.99	47.35	46.99	45.44
Gemma3n-EXAM^2 (Ours)	✔	✔	**64.31**	59.26	**60.14**	**61.24**
Random guess	✔	✔	27.18	24.97	25.02	25.72
Human ceiling (Ours)	✔	✔	91.59	84.08	65.66	80.44

Table 3: Average multilingual and multimodal evaluation results across the six languages over Speech, Sound, and Music. **Bold** and <u>underline</u> indicate the best and second best performance for each category among open-source models only. † shows cascaded input of transcript and/or caption. ✔ and ✓ are marked to differentiate the original audio waveform and textual input, respectively. ✘ marks the unexploited modality.

els to assess the impact of multimodal grounding and multilingual competence. Phi-4-multimodal-instruct variant[11], along with other models such as Qwen-2-audio-7B-instruct[12] Qwen-2.5-omni-4-bit-GPTQ[13] and Gemma-3n-E4B[14], are evaluated as open-source alternatives.

5.3 Results and Analysis

Table 3 summarizes the average performance on the EXAM^2 benchmark across three audio domains (speech, sound, and music) and six languages (DE, EN, ES, JA, MS, and ZH). Among closed-source models, GPT-4o-audio achieves the strongest overall performance with an average score of 73.83%, consistently leading in Speech, Sound, and Music. Focusing on open-source models, our proposed Gemma3n-EXAM^2 obtains the best overall average of 61.24% (with a significant improvement over Gemma3n, by paired t-test, $p < 0.05$), and achieves the highest scores in both Speech and Music with 64.31% and 60.14%, respectively. However, Gemma3n† (given the cascaded textual input) remains stronger in Sound, achieving 67.12% compared with 59.26% for Gemma3n-EXAM^2, indicating that our model is particularly effective for speech- and music-related understanding while leaving room for further improvement on sound-centric reasoning.

Meanwhile, Table 4 presents our results across all six languages. Overall, our proposed Gemma3n-EXAM^2 model, compared to baseline with the same modalities input, demonstrates competitive performance across all languages and audio categories, with a notable increase of 26.43% in the speech domain for German and +22.32% accuracy gain for all languages on average. A global improvement of 11.91% and 13.15% in the sound and music domain respectively is observed, as both sound and music benefit from the multimodal grounding provided by the visual answer choices. By utilizing our OmniLoRA approach on Gemma3n-E4B with our multilingual and multimodal training data, we can significantly boost the performance of the model across all languages and audio domains, demonstrating the high quality of our benchmark for improving multilingual and multimodal audio understanding.

Image modality helps audio understanding. For all languages, we notice a significant improvement by incorporating visual context into audio understanding, particularly for native multimodal settings. Models with joint image–audio capability frequently outperform their audio-only or cascaded counterparts on speech, environmental sound, and music understanding tasks. For instance, Qwen-2.5-omni exhibits a substantial performance gap between its audio+image configuration and audio-only variants (+15.55%, statistically significant across languages with paired t-test, $p < 0.01$), while the Gemma3n model shows slightly increased performance for Japanese, Malay, and Chinese. Sound category benefits the most from the visual information (+11.71% for Qwen-2.5-omni and +4.31% for Gemma3n among languages), which is expected as the sound-related questions often require more contextual information to be correctly answered. These findings suggest that visual grounding can provide complementary semantic cues for acoustic reasoning, especially for ambiguous audio events where contextual scene information disambiguates source identity,

[11] https://huggingface.co/microsoft/Phi-4-multimodal-instruct

[12] https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct

[13] https://huggingface.co/Qwen/Qwen2.5-Omni-7B

[14] https://huggingface.co/google/gemma-3n-E4B-it

Model	Input Modality		German		English		Spanish	
	Image	Audio	Speech / Sound / Music	Avg.	Speech / Sound / Music	Avg.	Speech / Sound / Music	Avg.
Closed-source								
GPT-4o-mini-audio	✘	✔	75.08 / 59.46 / 56.63	63.72	74.77 / 59.16 / 57.53	63.82	71.47 / 62.46 / 59.34	64.42
GPT-4o-audio	✘	✔	78.98 / 70.87 / 70.18	73.34	77.18 / 72.67 / 71.39	73.75	75.98 / 75.08 / 72.29	74.45
GPT-4o†	✔	✓	65.77 / 69.07 / 57.23	64.02	68.77 / 70.57 / 59.64	66.33	67.57 / 68.17 / 59.04	64.93
GPT-5-mini†	✘	✓	72.37 / 51.05 / 54.22	59.21	72.67 / 55.56 / 55.72	61.32	72.67 / 50.45 / 55.12	59.41
GPT-5-2†	✘	✓	71.17 / 48.05 / 51.51	56.91	72.67 / 51.95 / 50.00	58.21	69.07 / 54.05 / 51.51	58.21
Open-source								
Phi-4-multimodal-instruct	✘	✔	61.56 / 57.66 / 53.31	57.51	**65.17** / 65.57 / **64.16**	**64.97**	61.86 / 55.56 / 56.63	58.02
Phi-4-multimodal-instruct†	✘	✓	45.95 / 61.86 / **59.94**	55.92	50.75 / 65.47 / 63.25	59.82	49.55 / 64.56 / 56.93	57.01
Phi-4-multimodal-instruct†	✔	✓	51.65 / 63.66 / 44.28	53.20	54.95 / 62.76 / 50.90	56.20	50.75 / 59.76 / 44.88	51.80
Qwen-2-audio	✘	✔	28.23 / 34.83 / 39.46	34.17	38.74 / 41.74 / 47.89	42.79	35.14 / 32.43 / 43.37	36.98
Qwen-2.5-omni†	✘	✓	51.65 / 65.17 / 58.43	58.42	58.86 / 68.17 / 58.73	61.92	53.45 / 63.36 / 56.33	57.71
Qwen-2.5-omni†	✔	✓	54.65 / 62.46 / 58.73	58.61	58.56 / 67.87 / 59.04	61.82	53.15 / 60.06 / 55.42	56.21
Qwen-2.5-omni	✘	✔	41.44 / 51.35 / 34.94	42.58	39.34 / 48.95 / 30.42	39.57	40.54 / 48.35 / 31.33	40.07
Qwen-2.5-omni	✔	✔	51.05 / 61.26 / 57.83	56.71	55.86 / 66.97 / 58.73	60.52	51.65 / 61.86 / 55.42	56.31
Gemma3n†	✘	✓	56.76 / **69.67** / 52.41	59.61	60.06 / 66.67 / 56.93	61.22	58.86 / **68.47** / 53.01	60.11
Gemma3n†	✔	✓	55.86 / 69.07 / 53.61	59.51	58.86 / **70.57** / 55.42	61.62	56.16 / 65.17 / 55.12	58.82
Gemma3n	✔	✘	41.44 / 48.35 / 46.39	45.39	42.94 / 51.65 / 48.19	47.59	39.94 / 47.75 / 45.48	44.39
Gemma3n	✘	✔	41.74 / 45.05 / 47.29	44.69	47.75 / 43.84 / 51.81	47.80	47.45 / 41.44 / 48.19	45.69
Gemma3n	✔	✔	39.94 / 48.05 / 46.39	44.79	43.24 / 48.35 / 48.49	46.69	40.24 / 48.95 / 47.89	45.69
Gemma3n-EXAM^2 (Ours)	✔	✔	**66.37** / 62.76 / **59.94**	**63.02**	64.56 / 62.16 / 59.94	62.22	**64.26** / 58.86 / **61.14**	**61.42**

Model	Input Modality		Japanese		Malay		Chinese	
	Im	Au	Speech / Sound / Music	Avg.	Speech / Sound / Music	Avg.	Speech / Sound / Music	Avg.
Closed-source								
GPT-4o-mini-audio	✘	✔	74.47 / 62.16 / 56.02	64.22	71.17 / 60.06 / 56.33	62.52	75.08 / 62.76 / 56.93	64.92
GPT-4o-audio	✘	✔	79.28 / 74.17 / 68.07	73.84	78.68 / 73.27 / 69.88	73.94	77.48 / 72.97 / 70.48	73.64
GPT-4o†	✔	✓	66.97 / 68.47 / 55.42	63.62	63.36 / 69.07 / 55.72	62.72	67.57 / 69.67 / 55.72	64.32
GPT-5-mini†	✘	✓	71.77 / 48.65 / 53.92	58.11	69.97 / 51.65 / 54.82	58.81	71.47 / 50.15 / 51.96	57.86
GPT-5-2†	✘	✓	72.37 / 48.05 / 48.19	56.20	70.57 / 52.25 / 50.90	57.91	72.97 / 47.15 / 48.49	56.20
Open-source								
Phi-4-multimodal-instruct	✘	✔	59.16 / 51.05 / 51.81	54.01	51.05 / 44.14 / 49.70	48.30	60.66 / 52.85 / 54.52	56.01
Phi-4-multimodal-instruct†	✘	✓	48.95 / 59.46 / 54.22	54.21	42.04 / 54.05 / 49.01	48.37	45.05 / 59.46 / 53.92	52.81
Phi-4-multimodal-instruct†	✔	✓	50.15 / 58.56 / 42.17	50.29	46.25 / 57.06 / 39.16	47.49	54.05 / 57.66 / 43.07	51.59
Qwen-2-audio	✘	✔	29.73 / 26.73 / 34.34	30.27	36.04 / 29.73 / 31.93	32.57	37.84 / 40.54 / 45.18	41.19
Qwen-2.5-omni†	✘	✓	51.65 / 63.36 / 53.01	56.01	48.65 / 59.16 / 54.82	54.21	57.36 / **65.77** / **57.83**	**60.32**
Qwen-2.5-omni†	✔	✓	48.95 / 58.56 / 50.00	52.50	48.05 / 56.46 / 53.92	52.81	56.16 / 61.26 / 55.12	57.51
Qwen-2.5-omni	✘	✔	38.74 / 48.35 / 33.43	40.17	38.14 / 52.55 / 32.53	41.07	39.64 / 49.85 / 32.83	40.77
Qwen-2.5-omni	✔	✔	50.15 / 59.16 / 48.19	52.50	46.55 / 56.46 / 52.71	51.91	57.96 / 63.96 / 56.93	59.62
Gemma3n†	✘	✓	59.76 / **69.07** / 53.92	**60.92**	54.65 / **64.56** / 56.33	58.51	60.36 / 64.26 / 52.71	59.11
Gemma3n†	✔	✓	60.06 / 65.77 / 49.70	58.51	53.45 / 63.66 / 55.12	57.41	58.56 / 63.96 / 51.81	58.11
Gemma3n	✔	✘	43.24 / 46.85 / 47.29	45.79	41.44 / 47.45 / 47.59	45.49	43.84 / 50.15 / 44.88	46.29
Gemma3n	✘	✔	41.74 / 42.34 / 45.78	43.29	34.53 / 43.24 / 49.10	42.29	43.84 / 42.34 / 47.89	44.69
Gemma3n	✔	✔	43.24 / 45.05 / 46.08	44.79	42.04 / 45.95 / 47.89	45.29	43.24 / 47.75 / 45.18	45.39
Gemma3n-EXAM^2 (Ours)	✔	✔	**63.36** / 56.76 / **59.64**	59.92	**63.06** / 57.96 / **62.65**	**61.22**	**64.26** / 57.06 / 57.53	59.62

Table 4: Multilingual and multimodal evaluation results (accuracy in %) on the EXAM^2 test set. **Bold** represents the best performance, while underline indicates the second best performance for each category and each language within open-source models. † shows cascaded input of transcript and/or caption. We also mark ✔ and ✓ to differentiate the original audio waveform and textual input, respectively. ✘ marks the unexploited modality.

activity, or intent.

Visual confusion occurs on cascaded modes. Despite the advantages of multimodal inputs, we identify a recurring failure mode in cascaded systems that process modalities independently before language reasoning. Cascaded variants of Gemma3n show degraded performance when image inputs are introduced alongside audio, indicating that visual information can distract rather than assist the model's reasoning process. This phenomenon is particularly visible in models adapted from speech-specific cases, where image conditioning sometimes lowers average accuracy across sound and music tasks. We hypothesize that fusion in cascaded pipelines induces representational competition: visual embeddings can be weakly correlated with the target acoustic semantics. Such visual confusion highlights the limitation of late-fusion designs and motivates researchers with tighter cross-modal alignment and shared latent representations.

Western languages perform better than eastern languages. A clear geographic and linguistic disparity emerges across the multilingual benchmark. Western languages, including German, English, and Spanish, generally obtain higher average scores than East Asian languages such as Japanese and Chinese across most model families. The trend

is particularly pronounced for sound and music understanding subtasks, where performance degradation in Japanese and Chinese remains substantial even among frontier proprietary models. This discrepancy likely reflects uneven multilingual audio pretraining distributions, differences in phonetic structure, and the relative scarcity of culturally diverse non-speech acoustic supervision. Our findings suggest that multilingual multimodal competence remains strongly biased toward high-resource western language ecosystems.

Cross-Linguistic Evaluation. We also observe a notable domain similarity structure across languages. Speech and sound performance display strong positive alignment (Pearson correlation ≈ 0.80), implying that languages benefiting from robust spoken-audio understanding also tend to perform well on environmental sound tasks. Interestingly, Malay shows the strongest music understanding score (62.65%) among all languages, suggesting that music-related reasoning differs from that in other domains. Overall, the results reveal that multilingual audio understanding is not uniformly distributed across domains. Speech generalization transfers relatively well across languages, whereas sound and music understanding expose residual multilingual gaps, particularly for eastern languages.

Test Lang.	Train Lang.	Modality	Accuracy (%)	
			Speech / Sound / Music	Avg.
EN	ALL	Audio+Image	**64.6 / 62.2 / 59.9**	**62.2**
	EN	Audio+Image	61.9 / 58.0 / 56.6	58.8
	ALL	Audio	67.0 / 57.7 / 61.1	61.9
	ALL	Image	40.5 / 43.5 / 43.4	42.5
	-	Image	30.3 / 33.9 / 38.5	34.3
DE	ALL	Audio+Image	**66.4 / 62.8 / 59.9**	**63.0**
	EN	Audio+Image	48.9 / 49.2 / 45.8	47.8
	ALL	Audio	64.6 / 57.1 / 57.2	59.6
	ALL	Image	38.7 / 42.9 / 41.6	41.1
	-	Image	30.3 / 29.4 / 32.2	30.7
ES	ALL	Audio+Image	**64.3 / 58.9 / 61.1**	**61.4**
	EN	Audio+Image	45.8 / 50.2 / 48.6	48.2
	ALL	Audio	68.2 / 57.1 / 57.5	60.9
	ALL	Image	40.5 / 45.3 / 41.9	42.6
	-	Image	35.1 / 29.4 / 36.8	33.8
JA	ALL	Audio+Image	**63.4 / 56.8 / 59.6**	**59.9**
	EN	Audio+Image	49.5 / 43.8 / 43.1	45.5
	ALL	Audio	64.9 / 54.7 / 59.3	59.6
	ALL	Image	38.4 / 40.5 / 40.7	40.0
	-	Image	28.8 / 25.8 / 30.7	28.5
MS	ALL	Audio+Image	**63.1 / 58.0 / 62.7**	**61.2**
	EN	Audio+Image	51.1 / 47.1 / 42.5	46.9
	ALL	Audio	65.8 / 55.9 / 58.1	59.9
	ALL	Image	41.7 / 45.0 / 40.1	42.3
	-	Image	27.9 / 27.3 / 30.4	28.6
ZH	ALL	Audio+Image	**64.3 / 57.1 / 57.5**	**59.6**
	EN	Audio+Image	51.4 / 45.6 / 40.4	45.8
	ALL	Audio	67.0 / 53.8 / 54.5	58.4
	ALL	Image	38.1 / 40.2 / 38.3	38.9
	-	Image	29.2 / 32.1 / 35.2	31.9

Table 5: Ablation study for our Gemma3n-EXAM^2 on the EXAM^2-test by omitting different modalities on English-only and all six languages training data. - denotes the model is evaluated via purely generated images without training on any language data.

6 Ablation Study

6.1 Ablation on Modality

First, we investigate the contribution of input modality by comparing multimodal (audio+image), audio-only, and image-only variants of Gemma3n-EXAM^2 on the EXAM^2-test set across six languages. Overall, multimodal conditioning consistently provides the strongest performance. Averaged across all languages, multimodal module achieves 61.22%, outperforming audio-only (60.05%) and substantially exceeding image-only (41.23%). The improvement is particularly pronounced for environmental sound understanding, where multimodal inputs yield systematic gains over audio-only inference in every language, indicating that visual grounding supplies complementary contextual cues for disambiguating non-speech acoustic events.

Interestingly, the relative benefit of modality differs across domains. Speech understanding achieves the best performance under audio-only setting, showing higher scores than audio+image in most languages. In contrast, multimodal fusion model provides clearer advantages for sound and, to a lesser extent, music understanding, where contextual scene information can help infer semantic intent, source identity, or acoustic structure. The non-trivial performance of image-only models suggests that visual priors encode weak but exploitable correlations between scenes and expected acoustic events. These findings reveal a complementary modality relationship: audio provides the primary semantic signal, while visual context improves robustness and cross-domain generalization, particularly for non-speech audio understanding.

6.2 Ablation on Language

We further examine the effect of training language composition by comparing models trained on English-only (EN) data against models trained on all six languages (ALL) under identical multimodal settings. Multilingual training produces substantial and highly consistent improvements across all evaluation languages. Monolingual training leads to a degradation of 12.39 points (48.83% averaged across languages). This pattern is univer-

sal across all six languages.

Evidently, multilingual training benefits not only non-English languages but also English itself with a slight improvement of $+3.4$ points, suggesting multilingual exposure acts as a regularizer rather than introducing harmful language interference, improving representation learning even for the dominant training language. A domain-wise breakdown further reveals that multilingual learning particularly strengthens sound and music understanding. For instance, in the German evaluation, multilingual training improves sound and music accuracies ($+13.6$ and $+14.1$ gains respectively). Similar gains appear across Japanese, Malay, and Chinese. We explain that multilingual supervision enriches the diversity of acoustic-language alignment during training, leading to representations that generalize better across culturally and linguistically heterogeneous auditory phenomena.

6.3 Ablation on Purely Generated Images

To investigate the contribution of human quality checking, we conduct an ablation study comparing the performance of our model evaluated on human-checked images versus purely generated images without the 25.05% of human modified real screenshots. Table 5 shows this solely image-generated setting leads to a substantial performance degradation across all six languages. In contrast, when the corresponding human-validated images are used, the average accuracies of the Gemma3n reach 46.70% (EN), 44.79% (DE), 45.69% (ES), 44.79% (JA), 45.29% (MS), and 45.39% (ZH), respectively. The results also show that training provides a clear benefit. Comparing the trained image-only setting (ALL) with the untrained setting (-), training improves average accuracy across all languages. This consistent improvement across languages indicates that multilingual training substantially strengthens the model's ability to extract and utilize information from the generated visual modality and suggests that manual visual refinement improves the quality of the visual modality rather than introducing exploitable generation artifacts.

7 Conclusion

To summarize, we introduce EXAM^2, a comprehensive benchmark with in total $135,684$ multilingual and $22,614$ multimodal instances for evaluating audio understanding capabilities of LALMs, LLMs, MLLMs, and cascaded systems. We conduct extensive evaluations of open- and closed-source models across three audio domains (speech, sound, and music) and six languages (DE, EN, ES, JA, MS, and ZH). Results reveal that multimodal conditioning significantly enhances audio understanding, particularly for non-speech sound tasks, while multilingual training substantially improves performance across all languages, especially for eastern languages. Our proposed Gemma3n-EXAM^2 model, trained on our multilingual and multimodal training data, achieves competitive performance across all languages and audio categories, demonstrating the high quality of our benchmark for improving multilingual and multimodal audio understanding. Future work can explore more efficient training methods and a wider range of languages to further enhance the multilingual and multimodal capabilities for the EXAM^2 benchmark.

Limitations

There are several limitations to our current evaluation of the EXAM^2 benchmarks. In these benchmarks, we have prioritized accuracy over efficiency. Future evaluations should incorporate inference speed and deployment constraints to enable a more comprehensive assessment of model performance in real-world applications. Second, due to limited computational resources, we have only curated a subset of the AudioMCQ dataset to generate visual representations for training our model. Future work can explore the full dataset and consider more efficient training methods to leverage the entire dataset effectively. Additionally, our language coverage is limited to European and Asian languages, a wider range of languages such as African language families can be included to better assess the multilingual capabilities of models on the EXAM^2 benchmarks.

Ethics Statement

The authors are not purposely creating or using any visual data that contains personally identifiable information or offensive content. Generated images are solely used for training and evaluation purposes on the EXAM^2 benchmarks, and we have taken care to ensure that they do not contain any harmful or inappropriate content.

Acknowledgements

This research is supported by UDF01004676 and the National Research Foundation, Singapore un-

der its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-004), and by A*STAR under its RIE2025 Japan-Singapore Joint Call: Japan Science and Technology Agency (JST) and Agency for Science, Technology and Research (A*STAR) 2024 (Award R24I6IR136). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

Use of AI Tools

We have used GPT-5-mini for translating the original English questions and choices into other five languages, and we have used GPT-image-2 for generating the visual representations of the audio. The authors also acknowledge the use of ChatGPT for assistance with grammar, punctuation, and vocabulary refinement, as well as debugging tasks.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore.

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, and 1 others. 2025. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico.

Chen Chen, ZeYang Hu, Fengjiao Chen, Liya Ma, Jiaxing Liu, Xiaoyu Li, Ziwen Wang, Xuezhi Cao, and Xunliang Cai. 2025. Uno-bench: A unified benchmark for exploring the compositional law between uni-modal and omni-modal in omni models. *arXiv preprint arXiv:2510.18915*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and 1 others. 2026. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Advances in Neural Information Processing Systems*, 38:41819–41886.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Evuru, S Sakshi, Oriol Nieto, Ramani Duraiswami, Dinesh Manocha, and 1 others. 2024. Compa: Addressing the gap in compositional reasoning in audio-language models. In *International Conference on Learning Representations*, volume 2024, pages 15486–15511.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 976–980. IEEE.

Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi Liu, Zining Liang, and 1 others. 2025. Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models. *arXiv preprint arXiv:2509.21060*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, and 1 others. 2024. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024-2024 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 12136–12140. IEEE.

Laya Iyer, Angelina Wang, and Sanmi Koyejo. 2026. SCENEBench: An audio understanding benchmark grounded in assistive and industrial use cases. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7123–7137, Rabat, Morocco.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv preprint arXiv:2402.01831*.

Sonal Kumar, Simon Sedlácek, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plicka, Miroslav Hlavácek, William Fineas Ellingwood, Sathvik Udupa, Siyuan Hou, Allison Ferner, Sara Barahona, Cecilia Bolaños, Satish Rahi, Laura Herrera-Alarcón, Satvik Dixit, and 15 others. 2026. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 22688–22697. AAAI Press.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, and 1 others. 2025. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1140–1144. IEEE.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, and 15 others. 2025. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *CoRR*, abs/2505.13032.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2024. Mustango: Toward controllable text-to-music generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8293–8316, Mexico City, Mexico.

Zi Haur Pang, Xiaoxue Gao, Tatsuya Kawahara, and Nancy F Chen. 2026. Erm-minmaxgap: Benchmarking and mitigating gender bias in multilingual multimodal speech-llm emotion recognition. *arXiv preprint arXiv:2603.21050*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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

Qwen Team. 2026. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025a. AudioBench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316, Albuquerque, New Mexico.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025b. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*.

Shuting Wang, Jiejun Tan, Zhicheng Dou, and Ji-Rong Wen. 2025c. OmniEval: An omnidirectional and automatic RAG evaluation benchmark in financial domain. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5726–5751, Suzhou, China.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. *arXiv preprint arXiv:2408.01337*.

Valentin Werner. 2023. English and German pop song lyrics: Towards a contrastive textology. In *Journal for Language Technology and Computational Linguistics, Vol. 36 No. 1*, pages 1–20, unknown. German Society for Computational Lingustics and Language Technology.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2025. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, Vienna, Austria.

Zihan Zhao, Yiyang Jiang, Heyang Liu, Yu Wang, and Yanfeng Wang. 2024. Librisqa: A novel dataset and framework for spoken question answering with large language models. *IEEE Transactions on Artificial Intelligence*.

A Appendix

A.1 Dataset Details

A.1.1 Translation Guidelines

When translating the original English questions and choices into other five languages, we first use the GPT-5-mini to follow the instruction in Figure 3. The translated results are then carefully reviewed by native speakers of the target languages (or researchers who are fluent in those languages for over 5 years) to ensure the accuracy, fluency, and appropriateness of the translations. Any issues identified during the review process were addressed and corrected to ensure the highest quality of the translated questions and answer choices.

Below is our specific guideline to ensure the quality and consistency of the translations:

1. **Preserve Meaning**: The translation should accurately convey the meaning of the original English text, including the question and all answer choices.

2. **Maintain Format**: The structure of the question and answer choices should be preserved in the translation, including the numbering and formatting.

3. **Use Natural Language**: The translation should use natural and fluent language that is appropriate for the target language, avoiding literal translations that may sound awkward or unnatural.

4. **Cultural Relevance**: If the original question contains culturally specific references or examples, we may adapt them to be more relevant to the target language and culture while maintaining the original intent of the question.

5. **Consistency Across Languages**: We aimed to maintain consistency in the translation style and terminology across all languages to ensure that the questions are comparable and that the evaluation is fair across different linguistic contexts.

A.2 Model Overview

A.2.1 Multilingual Examples

Figure 7 and Figure 6 shows the multilingual translation validated by native speakers on the EXAM^2-test set. By not only checking the correctness and fluency of the translations, we also ensure that the translated answer in a consistent position with the original English answer. For instance, we manually removed the redundant indexes of the given options, especially for the German and Spanish, as GPT-5-mini tends to automatically add the inconsistent indexes in the translated choices, which may cause confusion for the model. Additionally, if the audio contains some specific words that are asked in the question, we keep those words unchanged in the translated questions and choices, as shown in Figure 6. This step aligns with our processed transcript and caption, as most of the audio contains English words, and we want to ensure the consistency between the content within languages.

Translate the following audio understanding/reasoning question, its choices, and its answer into {Language}.
Keep technical/music notation unchanged.
Keep the English words in the choices unchanged if they are asked which word appears first in the question or something similar.
Make sure the translated answer matches one of the translated choices.

Question:
{question}

Choices:
{option 1. } {option 2. } ...

Answer:
{answer}

Figure 3: Translation prompt of GPT-5-mini. {Language} is replaced by the target language (German, Spanish, Japanese, Malay, and Chinese) when applying the prompt.

A real-world and precise visual representation of the audio:
{option 1. }, {option 2. }, ...

Figure 4: Image generation prompt of GPT-image-2 and stable diffusion.

A.2.2 Visual Examples

We provide visual examples of generated images for all three audio categories (speech, sound, and music) in Figure 8 and Figure 9. For images in Figure 8a, Figure 8b, and Figure 8c, they are screenshots manually created by authors, as the original images generated by stable diffusion/GPT-image-2 are not highly relevant to the audio content. However, due to strong capability of visual understanding in environmental sounds, GPT-image-2 can generate more fine-grained images for sound understanding questions, which is particularly helpful in our EXAM^2-train, as shown in Figure 9.

You are an expert in multimodal audio understanding. Let us solve a multiple-choice question together.
Image 1 -> Choice 1
Image 2 -> Choice 2
Image 3 -> Choice 3
Image 4 -> Choice 4

Use ALL inputs:
- Question
- Choices
- Transcript
- Caption
- Images

Instructions:
- Select the SINGLE best answer.
- Return ONLY the exact choice text.
- Do not output the choice number.
- Do not explain.
- Do not output anything else.

Question:
{question}

Choices:
{choices_text}

Transcript:
{transcript}

Caption:
{caption}

Figure 5: Inference prompt for multimodal models including Phi4-multimodal-instruct, Qwen2.5-omni, and Gemma3n.


```
"audio_id": "./test-mini-audios/421fcbf8-7f60-4770-8923-512323e5efba.wav",
"question": "Which word appears first",
"choices": [
  "push-ups",
  "pull-ups"
],
"answer": "push-ups",
"clotho": "a shuffled while snickers cat children talks",
"transcript": "Push-ups, press-ups, pull-ups, sit-ups.",
"answer_id": "1",
"question_de": "Welches Wort erscheint zuerst?",
"choices_de": [
  "push-ups",
  "pull-ups"
],
"answer_de": "push-ups",
"question_es": "¿Qué palabra aparece primero?",
"choices_es": [
  "push-ups",
  "pull-ups"
],
"answer_es": "push-ups",
"question_ja": "どの単語が最初に現れますか？",
"choices_ja": [
  "push-ups",
  "pull-ups"
],
"answer_ja": "push-ups",
"question_ms": "Perkataan mana muncul terlebih dahulu",
"choices_ms": [
  "push-ups",
  "pull-ups"
],
"answer_ms": "push-ups",
"question_zh": "哪个单词先出现？",
"choices_zh": [
  "push-ups",
  "pull-ups"
],
"answer_zh": "push-ups"
```


Figure 6: Transcript and caption, multilingual translation examples in `json` screenshot on our EXAM^2.

Hyperparameter	Value
Data Type	bfloat16
Learning Rate	2e-4
Batch Size	4
Gradient Accumulation Steps	8
Training Epochs	6
Lora Alpha	32
Lora Dropout	0.05

Table 6: Hyperparameters used for supervised fine-tuning of Gemma3n.

A.3 Prompts

Figure 3 shows the prompt we used for translating the original English questions and choices into other five languages (DE, ES, JA, MS, and ZH) using GPT-5-mini. For simplicity, prompt to generate our image using GPT-image-2 and stable diffusion is shown in Figure 4.

When evaluating the performance of multimodal models, we use the same prompt as shown in Figure 5 for all multimodal models. Specifc modalities (e.g., image-only or audio-only) are controlled by removing the corresponding input from the prompt. For instance, for image-only evaluation, we only keep the question, choices, and images in the prompt, while for audio-only evaluation, we only keep the question, choices, and original audio in the prompt. Cascaded evaluation is conducted by applying the transcript / caption instead of the audio input in the prompt.

A.4 Experiment Details

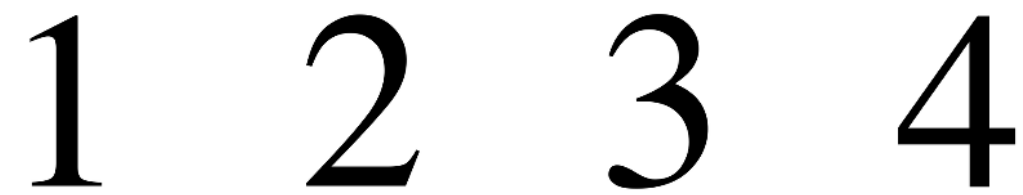
We give additional detailed hyperparameters used in our experiments in Table 6. The backbone model for fine-tuning is Gemma3n-E4B, and we use the OmniLoRA method for efficient fine-tuning.

```
"audio_id": "./test-mini-audios/69631267-f7ef-464e-8bc6-4f3e75e6fb6f.wav",
"question": "Based on the given audio, which of the following best describes the sound environment?",
"choices": [
  "A forest ambience with natural bird sounds only",
  "A live orchestra performing in a theater",
  "An outdoor scene with electric guitar music and birds in motion",
  "A crowded marketplace with street vendors"
],
"answer": "An outdoor scene with electric guitar music and birds in motion",
"question_de": "Basierend auf der gegebenen Audioaufnahme, welche der folgenden Beschreibungen trifft die Klangumgebung am besten?",
"choices_de": [
  "Eine Waldatmosphäre mit ausschließlich natürlichen Vogelstimmen",
  "Ein Live-Orchester, das in einem Theater spielt",
  "Eine Außenszene mit E-Gitarrenmusik und Vögeln in Bewegung",
  "Ein belebter Marktplatz mit Straßenverkäufern"
],
"answer_de": "Eine Außenszene mit E-Gitarrenmusik und Vögeln in Bewegung",
"question_es": "Basándose en el audio proporcionado, ¿cuál de las siguientes opciones describe mejor el entorno sonoro?",
"choices_es": [
  "Un ambiente forestal con sonidos naturales de aves únicamente",
  "Una orquesta en vivo actuando en un teatro",
  "Una escena al aire libre con música de guitarra eléctrica y aves en movimiento",
  "Un mercado concurrido con vendedores ambulantes"
],
"answer_es": "Una escena al aire libre con música de guitarra eléctrica y aves en movimiento",
"question_ja": "与えられたオーディオに基づいて、次のうちどれが音の環境を最もよく表していますか？",
"choices_ja": [
  "自然の鳥の鳴き声のみが聞こえる森林の音環境",
  "劇場で演奏しているライブオーケストラ",
  "エレクトリックギターの音楽と鳥が飛び回る屋外のシーン",
  "屋台のある混雑した市場"
],
"answer_ja": "エレクトリックギターの音楽と鳥が飛び回る屋外のシーン",
"question_ms": "Berdasarkan audio yang diberikan, yang manakah berikut paling tepat menggambarkan persekitaran bunyi?",
"choices_ms": [
  "Suasana hutan dengan bunyi burung semula jadi sahaja",
  "Orkestra langsung yang tampil di teater",
  "Adegan luar dengan muzik gitar elektrik dan burung yang bergerak",
  "Pasar yang sesak dengan peniaga jalanan"
],
"answer_ms": "Adegan luar dengan muzik gitar elektrik dan burung yang bergerak",
"question_zh": "根据所提供的音频，下列哪一项最能描述该声音环境？",
"choices_zh": [
  "只有自然鸟鸣的森林氛围",
  "在剧院演出的现场管弦乐团",
  "伴有电吉他音乐和活动中鸟类声音的户外场景",
  "有街头小贩的拥挤市场"
],
"answer_zh": "伴有电吉他音乐和活动中鸟类声音的户外场景"
```


Figure 7: Multilingual translation examples in `json` screenshot on our EXAM^2.



(a) Visual examples of speech category on EXAM^2 evaluation dataset. The generated images are from choices prompts (from left to right): `An hour; Two hours; Thirty minutes; Fifteen minutes.`



(b) Visual examples of sound category on EXAM^2 evaluation dataset. The generated images are from choices prompts (from left to right): `Once; Twice; Three times; Four times.`



(c) Visual examples of music category on EXAM^2 evaluation dataset. The generated images are from choices prompts (from left to right): `D#:min7/1; G#:min6(9,*1)/6; F#:maj7/1; C#:sus2(b7,*1)/b7.`

Figure 8: Visual examples manually created on EXAM^2 evaluation set.



(a) Visual examples of speech category on EXAM^2-Clotho training dataset. The generated images are from choices prompts (from left to right): `A concert host introducing a band; A sports commentator describing a match; An auction announcer addressing a crowd; A train station departure announcer.`



(b) Visual examples of sound category on EXAM^2-Clotho training dataset. The generated images are from choices prompts (from left to right): `The rain and ocean waves; The hail and mountain stream; The wind and forest animals; The thunder and city traffic.`



(c) Visual examples of music category on EXAM^2-Clotho training dataset. The generated images are from choices prompts (from left to right): `The violin and the gong; The xylophone and the gong; The piano and the gong; The xylophone and the cymbals.`

Figure 9: Visual examples generated by GPT-image-2 on EXAM^2-Clotho training set.