OmniACBench: A Benchmark for Evaluating Context-Grounded Acoustic Control in Omni-Modal Models

Anonymous ACL submission

Abstract

Most testbeds for omni-modal models assess multimodal understanding via textual outputs, leaving it unclear whether these models can properly *speak* their answers. To study this, we introduce OmniACBench, a benchmark for evaluating *context-grounded acoustic control* in omni-modal models. Given a spoken instruction, a text script, and an image, a model must read the script aloud with an appropriate tone and manner. OmniACBench comprises 3,559 verified instances covering six acoustic features: speech rate, phonation, pronunciation, emotion, global accent, and timbre. Extensive experiments on eight models reveal their limitations in the proposed setting, despite their strong performance on prior textual-output evaluations. Our analyses show that the main bottleneck lies not in processing individual modalities, but in integrating multimodal context for faithful speech generation. Moreover, we identify three common failure modes—weak direct control, failed implicit inference, and failed multimodal grounding—providing insights for developing models that can verbalize responses effectively.

1 Introduction

Multimodal Large Language Models (MLLMs) have rapidly evolved from bi-modal systems—such as text–vision (Alayrac et al., 2022; Li et al., 2023) and text–audio (Deshmukh et al., 2023; Tang et al., 2023)—to omni-modal architectures that jointly process text, vision, and audio as input (Han et al., 2024; Li et al., 2024a).[1] More recently, this trend has extended beyond inputs: modern omni-modal models can now generate speech responses as well as text (Xu et al., 2025a,b; Wang et al., 2025a; Tong et al., 2025). This shift marks a transition for omni-modal systems from multimodal understanding to generating responses in diverse modalities.

[1]**Omni-modal** refers to the text–vision–audio setting, while **multimodal** covers any combination of modalities.

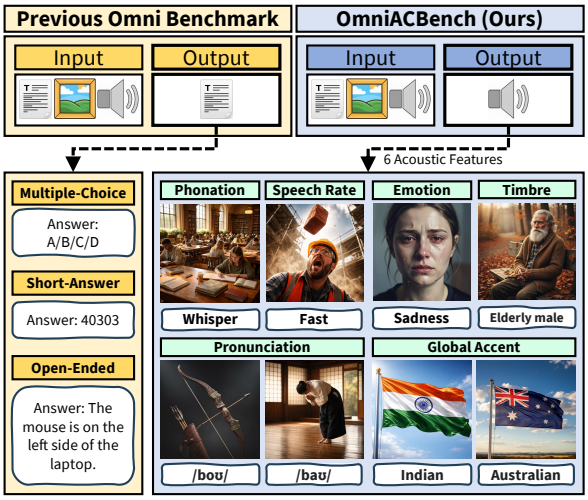

Figure 1: Comparison of prior omni-modal benchmarks and OmniACBench. Existing ones assess multimodal understanding via text outputs, whereas ours targets speech generation given text, vision, and speech inputs.

Alongside these advances, several benchmarks have been introduced to evaluate omni-modal models (Li et al., 2024b; Hong et al., 2025; Zhou et al., 2025; Kim et al., 2025a; Chen et al., 2026). Most of them focus on testing multimodal understanding, examining whether systems can interpret multimodal inputs and produce semantically correct textual answers. As a result, they provide valuable testbeds for measuring models' comprehension and reasoning abilities across modalities. However, as omni-modal models increasingly generate outputs in multiple formats, a key research question remains: **what should be considered and evaluated when responses are delivered in speech?**

Speech responses encode information not only in words but also in acoustic delivery, such as speaking rate, phonation, and other paralinguistic cues (Guyer et al., 2021). Therefore, speech-based evaluation should consider both whether the content is correct and whether the delivery is appropriate—a dimension with no direct counterpart in text-based

Benchmark	Input			Output	Target Features
	T	V	S		
OmniBench	✓	✓	✓	Text	Tri-Modal
WorldSense	✓	✓	✓	Text	Real World
Daily-Omni	✓	✓	✓	Text	Temporal
OMHBench	✓	✓	✓	Text	Multi Hop
OmniVideoBench	✓	✓	✓	Text	Diverse Video
UNO-Bench	✓	✓	✓	Text	Uni/Omni Link
AV-SpeakerBench	✓	✓	✓	Text	Speaker Centric
FutureOmni	✓	✓	✓	Text	Future Forecast
URO-Bench	✗	✗	✓	Speech	Em, Si, Re
VocalBench	✗	✗	✓	Speech	Em
S2S-Arena	✗	✗	✓	Speech	Em, Ti, SR Pr, Si, St
ParaS2SBench	✗	✗	✓	Speech	Em, Ti, Sa
VA-Eval-Viewing	✗	✓	✓	Speech	Visual QA
VA-Eval-Speaking	✗	✗	✓	Speech	Em, Ti
OmniACBench	✓	✓	✓	Speech	Em, Ti, SR, Pr, GA, Ph

Table 1: Comparing omni-modal and speech generation benchmarks by input/output modalities and target features. **Abbreviations: T**ext/**V**ision/**S**peech, **Em**otion, **Ti**mbre, **S**peech **R**ate, **Pr**onunciation, **G**lobal **A**ccent, **Ph**onation, **Si**nging, **Re**citation, **St**ress, **Sa**rcasm.

evaluation. The challenge becomes especially pronounced in omni-modal settings, where appropriate acoustic realization must be inferred from signals across different modalities. For example, a conversation in a library may call for whisper-like phonation, an emergency situation for a faster speaking rate, and a sorrowful scene for a sad vocal expression. We refer to the ability to generate acoustically appropriate speech given multimodal context as **context-grounded acoustic control**.

To test this aspect, we propose **OmniACBench**, a benchmark for context-grounded **a**coustic **c**ontrol in **omni**-modal models. Given a spoken instruction, a text script, and an image, the model must generate speech that faithfully reads the script while realizing an acoustic delivery consistent with the combined multimodal context. OmniACBench covers six acoustic features and evaluates both measurable and abstract properties of speech output. Figure 1 highlights the key differences between prior omni-modal datasets and OmniACBench.[2]

We evaluate eight models on OmniACBench and find that current systems perform poorly overall, even when they achieve strong results on text-based evaluations. Our analysis further reveals that this gap does not stem simply from failures in processing a specific modality, but from the difficulty of generating speech that preserves the target content while acoustically reflecting multimodal context.

[2]OmniACBench and its code will be publicly released.

2 Related Work

2.1 Omni-Modal Benchmarks

Several benchmarks have been introduced to assess omni-modal capabilities (see Table 1). OmniBench (Li et al., 2024b), WorldSense (Hong et al., 2025), Daily-Omni (Zhou et al., 2025), OmniVideoBench (Li et al., 2025a), and AV-SpeakerBench (Nguyen et al., 2025) examine how well systems understand inputs spanning text, vision, and audio. OMHBench (Kim et al., 2025a) further addresses modality shortcut issues by introducing omni-modal multi-hop reasoning, while UNO-Bench (Chen et al., 2025) shows that omni-modal capability is jointly determined by underlying uni-modal abilities. FutureOmni (Chen et al., 2026) focuses on forecasting future events grounded in audio-visual context. Most existing benchmarks study multi-modal understanding through text outputs, whereas OmniACBench shifts the focus to speech-based evaluation, assessing both semantic fidelity and context-grounded acoustic control.

2.2 Speech Generation Benchmarks

Recent benchmarks for speech generation have moved beyond textual correctness to evaluate how spoken responses are delivered. URO-Bench (Yan et al., 2025) and VocalBench (Liu et al., 2025) assess broad speech-interaction abilities, including acoustic and paralinguistic phenomena. S2S-Arena (Jiang et al., 2025) and ParaS2SBench (Yang et al., 2025b) focus more directly on speech-to-speech instruction following, examining whether models produce appropriate content and speaking style under spoken cues. VoiceAssistant-Eval (Wang et al., 2025b) further expands this line to voice-assistant scenarios spanning listening, speaking, and viewing. However, these benchmarks primarily evaluate speech-centric interaction or general spoken assistant behavior, whereas OmniACBench focuses on context-grounded acoustic control in speech generation under joint text, vision, and speech inputs.

3 OmniACBench

We present **OmniACBench**, a benchmark for evaluating context-grounded acoustic control in omni-modal models. Each instance in the dataset comprises a text script, a spoken instruction, and an image, requiring the model to read the script aloud with appropriate acoustic realization. The three

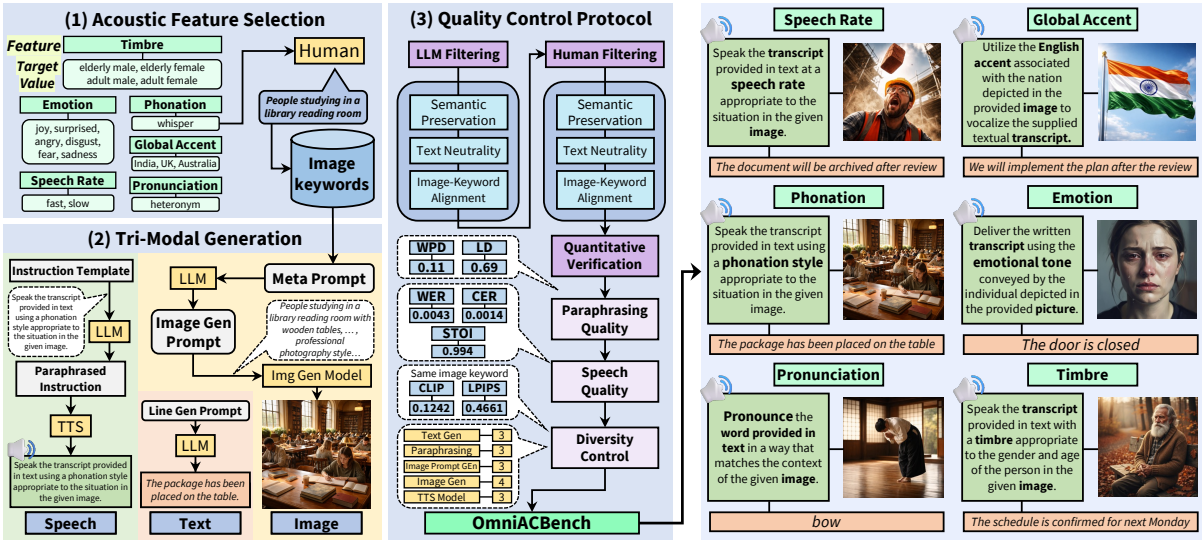

Figure 2: Construction pipeline of OmniACBench with representative examples for each acoustic feature. (1) **Acoustic Feature Selection** defines the target acoustic features and associated image keywords. (2) **Tri-Modal Generation** constructs each instance from a neutral text script, a spoken control signal, and a generated image. (3) **Quality Control Protocol** applies filtering and quantitative verification to ensure data quality and diversity.

input modalities play distinct roles: the *text* defines the linguistic content, the *spoken instruction* specifies the acoustic aspect to control, and the *image* provides contextual cues. By separating content, control, and contextual grounding across modalities, OmniACBench evaluates *context-grounded acoustic control* rather than treating the task as prosodic rendering or explicit style conditioning.

The construction of OmniACBench follows a three-stage pipeline: (1) **Acoustic Feature Selection**, (2) **Tri-Modal Generation**, and (3) **Quality Control Protocol**, as illustrated in Figure 2.

3.1 Acoustic Feature Selection

In the first stage, we select acoustic features and define their target values, each representing a specific realization (e.g., *fast* for Speech Rate or *angry* for Emotion). We choose these features based on two complementary criteria.

First, we consider **multimodal groundability**: each feature should admit natural visual grounding, so that images can provide meaningful clues for the intended target value (e.g., an emergency scene implicitly calling for a faster speech rate). Second, we consider **evaluation diversity**: the selected features should span different forms of acoustic evaluation. Some features correspond to measurable attributes that can be assessed with explicit objective metrics (e.g., Speech Rate measured in words per minute), whereas others are abstract and cannot be reduced to a single objective metric (e.g., Emotion). Including both types enables a more comprehensive evaluation of acoustic control.

Based on these criteria, we select six acoustic features: **Speech Rate**, **Phonation**, **Pronunciation**, **Emotion**, **Global Accent**, and **Timbre**. Each feature is associated with a predefined set of target values, and each benchmark instance is defined by a **feature–value** pair (e.g., Emotion–angry, Speech Rate–fast). For each target value, human annotators manually curate a set of image keywords describing visual scenes or concepts naturally associated with it, such as "people studying in a library reading room" for *Phonation–whisper* and "a boiling-over pot" for *Speech Rate–fast*.

3.2 Tri-Modal Generation

In the second phase, we construct the three modalities of each OmniACBench instance—*text*, *speech*, and *image*—from a given feature–value pair. All instances are synthetically generated using generative models, enabling scalable and controllable construction. This design also supports fine-grained analysis of model behavior (Sections 5.1 and 5.2).

Text The text modality provides the script that the model is required to read aloud. To prevent shortcut learning, the script is designed to remain neutral and not directly encode the intended acoustic characteristic. We thus generate scripts using an LLM, explicitly prompting it to produce content that is neutral with respect to attributes such as

emotion, nationality, gender, and age.

Speech The speech modality serves as a control signal that specifies which acoustic dimension should govern delivery. The model must use this signal to determine which acoustic property to infer from the image and realize in speech while reading the text. For each acoustic feature, we first design an instruction template and then use an LLM to generate instance-level paraphrases that increase linguistic diversity while preserving semantic intent. The resulting paraphrases are then synthesized into speech using a text-to-speech (TTS) model.

Image The image modality provides situational cues for inferring the target value within the acoustic dimension specified by the spoken instruction. Directly using the image keywords from the previous stage as prompts would limit both diversity and scalability. To address this, we design a meta-prompt that instructs an LLM to expand each keyword into an image-generation prompt in which the keyword remains the central visual focus while additional scene attributes are introduced. Specifically, the LLM is guided to produce a prompt consisting of 5 to 8 comma-separated visual elements. These prompts are then used by image generation models to synthesize the final images.

3.3 Quality Control Protocol

We apply a quality control protocol with two components: **Data Filtering** and **Quantitative Verification**. These procedures aim to detect and remove potential negative artifacts that may arise during the synthetic generation of OmniACBench.[3]

3.3.1 Data Filtering

We use three filtering strategies (Semantic Preservation, Text Neutrality, and Image–Keyword Alignment) applied at different stages of the pipeline. We first perform LLM-based filtering, followed by human verification using the same methods to double-check the results. **Semantic Preservation** tests whether paraphrased spoken instructions preserve the intent of the original sentence. **Text Neutrality** removes scripts whose content alone reveals the target acoustic value, which could otherwise enable shortcut learning. **Image–Keyword Alignment** checks whether generated images remain semantically consistent with their original keywords.

Starting from 3,640 generated instances, 3,586 remain after LLM-based filtering and 3,559 after human verification, yielding a final retention rate of 97.78%. The benchmark spans six acoustic features with roughly 600 instances each, indicating a balanced feature distribution. Detailed statistics are summarized in Table 5 of the Appendix. Figure 15 presents concrete examples from OmniACBench.

3.3.2 Quantitative Verification

To further validate the benchmark quality, we conduct quantitative verification, with results summarized in Table 6 in the Appendix.

Paraphrasing Quality We evaluate the linguistic diversity of paraphrased spoken instructions using Word Position Deviation (WPD) and Lexical Deviation (LD) (Liu et al., 2022). OmniACBench achieves WPD/LD scores of 0.11/0.69, compared to 0.12/0.42 on MRPC (Dolan and Brockett, 2005) and 0.07/0.13 on PAWS (Zhang et al., 2019), two standard paraphrasing datasets. These results suggest sufficient linguistic diversity in the benchmark.

Speech Quality We evaluate synthesized speech using WER (↓) and CER(↓) computed by Whisper-large-v3 (Radford et al., 2022) to measure transcription fidelity, and STOI (↑) to gauge intelligibility (Kumar et al., 2023). The resulting scores (WER 0.004, CER 0.001, and STOI 0.994) indicate near-perfect transcription fidelity and intelligibility.

Diversity Control First, we use meta-prompting rather than naïve keyword prompting. Its effectiveness is verified by measuring the average pairwise CLIP embedding cosine distance (Radford et al., 2021) and LPIPS (Zhang et al., 2018) within each keyword group. Meta-prompting yields higher values on both metrics (CLIP 0.1242 vs. 0.0671; LPIPS 0.4661 vs. 0.3733), indicating greater visual variation than keyword-only prompting.

Second, to reduce model-specific bias, we maintain diverse model pools for each generation stage—three for text generation, three for paraphrasing, three for image prompt generation, four for image generation, and three for TTS. For every generation call within the pipeline, we randomly sample one model from the corresponding pool. LLM-based filtering is performed using a separate model. The models used are listed in Table 7 in the Appendix.

4 Experiments

4.1 Evaluation Methods and Metrics

Semantic Fidelity We assess whether the generated speech preserves the target text. To do so,

[3]Refer to Appendix B for full details of our protocol.

Models	Semantic	Speech Rate	Pron.	Phonation	Emotion	Global Accent	Timbre
	WER ↓	ΔWPM ↑	PER ↓	VFR@0.3 ↑ (%)	Emo-Acc ↑ (%)	GA-Acc ↑ (%)	Tim-Acc ↑ (%)
Reference Scores	0.05	65.87	1.21	96.78	89.43	97.29	96.67
MiniCPM-o 4.5	**1.04**	**6.42**	**5.46**	**1.69**	**21.44**	**39.34**	24.66
InteractiveOmni 8B	1.23	-0.73	6.46	0.00	14.57	33.70	24.33
InteractiveOmni 4B	1.31	0.45	7.97	0.00	15.81	33.33	24.83
Qwen3-Omni 30B	2.14	-1.81	7.40	0.00	17.09	31.33	25.17
Qwen2.5-Omni 7B	4.15	0.76	10.27	0.85	19.10	28.96	24.66
Qwen2.5-Omni 3B	4.51	0.91	13.47	0.00	16.58	28.71	24.66
Uni-MoE-2.0-Omni	5.21	-2.82	9.27	1.69	16.75	36.25	25.17
MGM-Omni 7B	5.96	-0.60	21.97	0.00	16.42	36.61	**25.34**

Table 2: Main results on OmniACBench. Semantic denotes semantic fidelity, and Pron. denotes pronunciation. WER measures semantic fidelity; ΔWPM, PER, and VFR@0.3 evaluate speech rate, pronunciation, and phonation; Emo-Acc, GA-Acc, and Tim-Acc assess emotion, global accent, and timbre. The *Reference Scores* denote those from human or trained evaluators. Random baseline performance for the last three features is 16.7%, 33.3%, and 25.0%, respectively. Best in **bold**, second-best underlined.

we transcribe the generated speech using Whisper-large-v3 (Radford et al., 2022) and compute Word Error Rate (WER) against the reference script.

Measurable Acoustic Features **Speech Rate** is measured in Words Per Minute (WPM). To account for differences in default speaking speed across models, we report ΔWPM, defined as the difference between the average WPM of instances assigned the target values *fast* and *slow* (i.e., $\Delta\text{WPM} = \overline{\text{WPM}}_{\text{fast}} - \overline{\text{WPM}}_{\text{slow}}$). A larger ΔWPM indicates a clearer distinction between the two target values. **Pronunciation** is evaluated using Phoneme Error Rate (PER). We extract phoneme sequences from generated speech using POWSM (Li et al., 2025b) and compare them with the reference phoneme sequence. **Phonation** is evaluated via whisper detection. Because whispered speech exhibits limited vocal-fold vibration and thus little or no fundamental frequency (F_0) (Zhao and Lin, 2016; Gudepu et al., 2020), we compute the Voiced Frame Ratio (VFR), i.e., the proportion of frames with detected F_0, and classify an utterance as whisper-like if VFR ≤ 0.3. We validate this threshold on the Expresso dataset (Nguyen et al., 2023), where 90.8% of whisper-style and only 0.2% of normal-style utterances satisfy the criterion. We report the resulting detection rate as VFR@0.3.

Abstract Acoustic Features Following prior work that evaluates using model-based evaluators (Yan et al., 2025; Liu et al., 2025; Wang et al., 2025b; Yang et al., 2025a), we assess three features—**Emotion**, **Global Accent**, and **Timbre**—using task-specific evaluators trained for our label space. For each feature, we collect speech samples from diverse sources (11k for Emotion, 12k for Global Accent, and 12k for Timbre) and train a WavLM-Large-based classifier (Chen et al., 2022a); the details are provided in Appendix C. On held-out test sets, the evaluators achieve accuracies of 89.43%, 97.29%, and 96.67% for Emotion, Global Accent, and Timbre, respectively. We use these evaluators to score model-generated speech, reported as Emo-Acc, GA-Acc, and Tim-Acc.

4.2 Experimental Setup

We evaluate eight omni-modal models—MiniCPM-o 4.5 (OpenBMB, 2026), InteractiveOmni (8B and 4B) (Tong et al., 2025), Qwen3-Omni 30B (Xu et al., 2025b), Qwen2.5-Omni (7B and 3B) (Xu et al., 2025a), Uni-MoE-2.0-Omni (Li et al., 2025c), and MGM-Omni 7B (Wang et al., 2025a)—that take text, image, and speech inputs and generate speech outputs. Table 2 shows their performance on OmniACBench. We also report the **Reference Scores** for better interpretation: semantic fidelity and measurable features are computed from five human annotators, while abstract features are measured by held-out evaluator accuracy.

4.3 Main Results

Overall Trends The performance remains limited across models, highlighting the difficulty of OmniACBench for current omni-modal models. Even models such as Qwen3-Omni 30B, which perform strongly on prior omni-modal benchmarks (Kim et al., 2025a; Li et al., 2025a; Chen et al., 2025; Nguyen et al., 2025; Chen et al., 2026), show weak performance on OmniACBench. MiniCPM-o 4.5 achieves relatively stronger results on most met-

rics, yet remains far below the Reference Scores. These results suggest that OmniACBench captures capabilities that have not been sufficiently examined in existing benchmarks.

Semantic Fidelity We first examine whether models faithfully reproduce the target script during speech generation. Compared with the Reference Scores, all models exhibit substantially higher WER, indicating that semantic fidelity itself becomes challenging when speech generation is conditioned on multimodal inputs. That is, even before acoustic control is considered, models often fail to preserve the target text.

Measurable Acoustic Features Performance on measurable acoustic features—Speech Rate, Pronunciation, and Phonation—remains suboptimal for nearly all models. Most ΔWPM values stay close to zero, suggesting little ability to modulate speaking rate even when the context implies a clear difference in tempo. Likewise, high PER scores and near-zero VFR@0.3 values indicate that models rarely realize fine-grained pronunciation control or whisper-like phonation. Overall, these results show that mapping contextual cues to precise, objectively measurable acoustic variation remains a major challenge for current omni-modal models.

Abstract Acoustic Features Compared with measurable features, abstract features appear somewhat easier to model, though overall performance remains limited. MiniCPM-o 4.5 achieves clear gains over the random baselines on Emotion and Global Accent, suggesting that it can partially reflect contextually implied speaker attributes in generated speech. However, most other models remain close to chance on these features, and timbre control is weak for all models. In summary, current omni-modal models still struggle to ground contextual information into controllable acoustic attributes, even for high-level properties.

5 Analysis

In this section, we investigate why current omni-modal models perform poorly on OmniACBench. We first explore whether this difficulty arises from insufficient processing of specific modalities. We then perform controlled input decomposition to identify failure points and characterize the resulting failure modes. Finally, we conduct an additional linear probing analysis to examine whether multimodal context remains decodable up to the speech generation stage.

Models	Script-Only	Spoken Conditioning	Visual Cue Inference
	WER ↓	WER ↓	Acc ↑
MiniCPM-o 4.5	0.12	0.12	94.75
InteractiveOmni 8B	0.12	0.13	95.56
InteractiveOmni 4B	0.10	0.11	92.88
Qwen3-Omni 30B	0.11	0.13	97.50
Qwen2.5-Omni 7B	0.13	0.13	94.21
Qwen2.5-Omni 3B	0.14	0.15	92.19
Uni-MoE-2.0-Omni	0.12	0.12	92.59
MGM-Omni 7B	0.09	0.09	92.11

Table 3: Diagnostic results for three component capabilities: script-only speech generation, spoken-instruction conditioning, and visual acoustic cue inference. Strong performance across these controlled settings suggests that weak performance on OmniACBench is not explained by missing component abilities alone.

5.1 Fundamental Capability Assessment

To interpret whether poor performance on OmniACBench reflects limitations in fundamental, modality-specific skills, we conduct three diagnostic ablation studies isolating core components: (1) script-only speech generation, (2) spoken-instruction conditioning, and (3) visual acoustic cue inference. Table 3 reports the results.

Script-Only Speech Generation We first test whether models can reliably read a provided script without multimodal grounding or acoustic control. In this setting, models receive only the target script from OmniACBench and generate speech. WER is uniformly low across models, suggesting that literal script reading is not the primary bottleneck.

Spoken-Instruction Conditioning In this ablation, models receive a spoken instruction together with the target script and are asked to generate the script in speech. WER remains nearly unchanged from the script-only condition, suggesting that spoken-instruction conditioning does not substantially impair content reproduction.

Visual Acoustic Cue Inference Finally, we examine whether models can infer target acoustic values from visual context. To isolate this capability from speech generation, we reformulate the problem as a multiple-choice task. Given an image associated with a feature, the model must select the target value implied by the image (e.g., *Select the appropriate speech rate associated with the image provided: (1) fast (2) slow*). Models achieve high

Models	Em. (%)		GA. (%)		Ti. (%)	
	E	H	E	H	E	H
MiniCPM-o 4.5 (Ori.)	21.7	23.0	40.0	38.7	27.5	26.0
MiniCPM-o 4.5 (Ora.)	33.3	35.7	70.0	72.7	25.0	24.0
Qwen3-Omni 30B (Ori.)	18.3	16.7	33.3	30.7	25.0	24.5
Qwen3-Omni 30B (Ora.)	23.3	22.0	33.3	34.0	25.0	26.0

Table 4: Human validation on a class-balanced subset of abstract features. E/H denote evaluator/human accuracy. **Abbreviations: Em**otion, **G**lobal **A**ccent, **Ti**mbre, **Ori**ginal, **Ora**cle.

accuracy, indicating that they can recover acoustic cues from visual scenes reliably in isolation.

Overall, the outcomes imply that the difficulty of OmniACBench cannot be explained solely by failures in script reading, spoken-instruction conditioning, or visual cue inference. Instead, the main challenge potentially lies in integrating these components into context-grounded speech generation.

5.2 Controlled Input Decomposition

The original task in OmniACBench requires a model to preserve the target script, identify the relevant acoustic dimension from the spoken instruction, infer its intended value from visual context, and realize it in generated speech. To pinpoint where this process fails, we perform controlled input decomposition, progressively replacing non-text inputs with textual surrogates.

Starting from the **Original** setting (spoken instruction, image, and text script), we introduce four simplified variants: **S-to-T**, **I-to-T**, **All-to-T**, and **Oracle**. In **S-to-T**, the spoken instruction is replaced with its textual counterpart. In **I-to-T**, the image is replaced with descriptive keywords. In **All-to-T**, all inputs are textualized, removing multimodal integration while still requiring the model to infer the target acoustic value from contextual description. Finally, in **Oracle**, the target acoustic value is explicitly specified in the instruction (e.g., *"Speak the provided Target Script Text as fast as possible."*).[4] This decomposition is enabled by the synthetic construction of OmniACBench, as the control variables used during data generation can be directly reused as experimental conditions.

We further assess evaluator–human agreement on model-generated speech using a class-balanced subset of Emotion, Global Accent, and Timbre under two conditions: Original and Oracle. The subset comprises 130 benchmark instances (60/30/40 for Emotion/Global Accent/Timbre), yielding 520 generated clips in total across two models and two conditions, with five human annotations per clip. As shown in Table 4, evaluator accuracies closely track majority-vote human accuracies across features, models, and conditions, supporting their use as scalable proxies in subsequent experiments.

[4]Figure 16 illustrates the gap between All-to-T and Oracle.

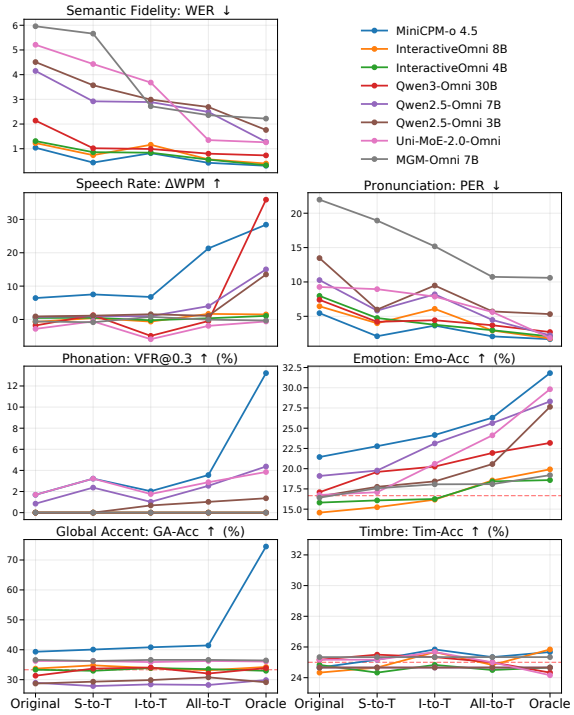

Figure 3: Results of Controlled Input Decomposition across all evaluation metrics. Starting from the Original setting, inputs are progressively textualized through S-to-T, I-to-T, and All-to-T, while Oracle explicitly specifies the target acoustic value. Dashed red lines in Emo-Acc, GA-Acc, and Tim-Acc indicate random baselines.

Overall Trends Figure 3 shows that performance generally improves as the task becomes more explicit and approaches the Oracle condition. However, both the magnitude and the shape of this improvement vary substantially across features and models, suggesting that the observed weakness cannot be reduced to a single bottleneck.

Failure Type I: Lack of Direct Acoustic Control For some features, performance remains at chance level even under Oracle conditions, indicating that the model cannot reliably realize the requested acoustic value even when no contextual inference is required. Timbre is the clearest example: all models remain near random across all conditions. Global Accent and Phonation show similar behavior for most models, with MiniCPM-

o 4.5 as the main exception. These results suggest that certain acoustic properties remain nearly uncontrollable in current omni-modal models.

Failure Type II: Failure of Implicit Acoustic Inference A second failure mode occurs when models can execute acoustic control once the target value is explicitly provided, but fail when it must be inferred from context. In such cases, performance increases sharply in Oracle while remaining at chance through All-to-T. For example, Speech Rate of Qwen3-Omni 30B, Qwen2.5-Omni 7B, and Qwen2.5-Omni 3B improves substantially when given explicit instructions, yet remains close to random when the target acoustic value must be inferred from context. A similar pattern appears for Phonation and Global Accent in MiniCPM-o 4.5. This suggests that direct acoustic control and implicit acoustic inference are separable capabilities, and that many models exhibit the former without reliably achieving the latter.

Failure Type III: Failure in Multimodal Acoustic Grounding The most challenging failure mode arises when models can infer the target acoustic value from textualized context but fail once the same information is distributed across modalities. MiniCPM-o 4.5 exhibits this pattern for Speech Rate. Its performance remains relatively strong in the All-to-T condition, indicating successful text-based inference, but drops in S-to-T and I-to-T. This suggests that multimodal acoustic grounding—linking contextually inferred acoustic intent to the actual acoustic realization of speech—remains a distinct and unresolved challenge, even when text-based inference and explicit control are available.

Emotion as a Relative Exception Emotion differs from the other features. Most models remain above chance across all conditions and improve steadily from Original to Oracle, suggesting that emotional cues are relatively easier to recover from context. At the same time, this contrast indicates that current omni-modal models capture emotional cues more reliably than broader acoustic properties.

5.3 Context Flow to Speech Generation

To better understand the performance gap on OmniACBench, we compare MiniCPM-o 4.5, the best-performing model on our benchmark, with Qwen3-Omni 30B, which performs strongly on prior benchmarks but less well in this setting. Using layer-wise hidden states, we train linear probes to predict the intended acoustic value, and use balanced accuracy to examine whether multimodal context remains linearly decodable as it propagates through each model. Figure 4 shows that MiniCPM-o 4.5 retains context-relevant acoustic information throughout the LLM backbone and into the TTS decoder, whereas Qwen3-Omni 30B drops from high decodability in the Thinker to chance in the Talker. This comparison points to a possible architectural advantage of models with tighter hidden-state integration across modality-specific components and the LLM, as in MiniCPM-o 4.5, over more decoupled Thinker–Talker designs such as Qwen3-Omni 30B for context-grounded acoustic control.

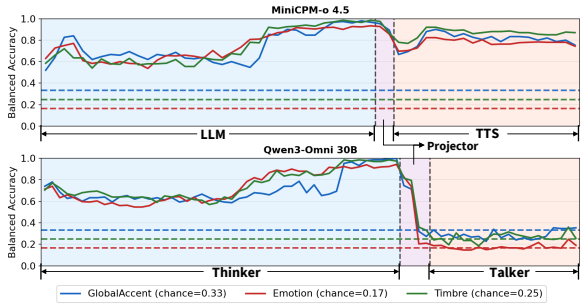

Figure 4: Linear probing of context-relevant information across model layers. MiniCPM-o 4.5 preserves decodable context into the TTS decoder, whereas Qwen3-Omni 30B drops to near chance in the Talker.

6 Conclusion

We proposed OmniACBench to evaluate a capability that has been largely overlooked in prior omni-modal evaluation: generating speech responses whose acoustic delivery correctly reflects multimodal context. Experiments on the benchmark reveal that strong performance on existing omni-modal benchmarks based on textual outputs does not readily transfer to this speech-oriented setting. The performance gap is not simply due to insufficient ability to perform individual elementary operations. Instead, our analyses point to three sources of weakness: limited direct control over certain acoustic attributes, challenges in inferring implicit targets from context, and unstable grounding under distributed multimodal information. Comparative probing further suggests that stronger acoustic control is associated with better retention of context-relevant information up to the speech generation stage. We hope OmniACBench will serve as a foundation for advancing context-grounded speech generation in omni-modal models.

Limitations

First, speech generation may involve joint control of multiple acoustic features (e.g., a fast speech rate and angry emotion), whereas each instance in OmniACBench evaluates only a single target acoustic feature. As an initial testbed, we leave this compositional setting to future work, especially given that current omni-modal models still struggle even in the single-feature setting.

Second, OmniACBench focuses on spoken instructions as the audio input, excluding non-speech signals such as environmental sounds or background music that may also provide contextual cues for acoustic delivery. Incorporating such audio contexts is left for future work.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Anthropic. 2025. Claude models overview. https://docs.anthropic.com/en/docs/about-claude/models/overview. Accessed: 2026-03-11.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

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*.

Qian Chen, Jinlan Fu, Changsong Li, See-Kiong Ng, and Xipeng Qiu. 2026. Futureomni: Evaluating future forecasting from omni-modal context for multimodal llms. *arXiv preprint arXiv:2601.13836*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022a. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Xueyuan Chen, Qiaochu Huang, Xixin Wu, Zhiyong Wu, and Helen Meng. 2022b. Hilvoice: Human-in-the-loop style selection for elder-facing speech synthesis. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 86–90. IEEE.

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.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

DTU54DL. common-accent. https://huggingface.co/datasets/DTU54DL/common-accent. Hugging Face dataset card, accessed March 13, 2026.

Kate Dupuis and M. Kathleen Pichora-Fuller. 2011. Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set. *Canadian Acoustics*, 39(3):182–183.

ElevenLabs. 2026a. Eleven flash v2.5 (model documentation). https://elevenlabs.io/docs/overview/models#flash-v25. Accessed: 2026-03-11.

ElevenLabs. 2026b. Eleven multilingual v2 (model documentation). https://elevenlabs.io/docs/overview/models#multilingual-v2. Accessed: 2026-03-11.

ElevenLabs. 2026c. Eleven turbo v2.5 (model documentation). https://elevenlabs.io/docs/overview/models#turbo-v25. Accessed: 2026-03-11.

Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michał Lis, Eshan Singhal, Nikola Mrkšić, Tsung-Hsien Wen, and Ivan Vulić. 2021. Multilingual and cross-lingual intent detection from spoken data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7468–7475.

Google. 2025a. Gemini 2.5 flash image model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf. Accessed: 2026-03-11.

Google. 2025b. Gemini 3 flash model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf. Accessed: 2026-03-01.

Google. 2025c. Gemini 3 pro image generation model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Image-Model-Card.pdf. Accessed: 2026-03-11.

Prithvi R.R. Gudepu, Gowtham P. Vadisetti, Abhishek Niranjan, Kinnera Saranu, Raghava Sarma, M. Ali Basha Shaik, and Periyasamy Paramasivam. 2020. Whisper Augmented End-to-End/Hybrid Speech Recognition System — CycleGAN Approach. In *Interspeech 2020*, pages 2302–2306.

Joshua J Guyer, Pablo Briñol, Thomas I Vaughan-Johnston, Leandre R Fabrigar, Lorena Moreno, and Richard E Petty. 2021. Paralinguistic features communicated through voice can affect appraisals of confidence and evaluative judgments. *Journal of nonverbal behavior*, 45(4):479–504.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2024. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26584–26595.

Sanaul Haq and Philip JB Jackson. 2011. Multimodal emotion recognition. In *Machine audition: principles, algorithms and systems*, pages 398–423. IGI Global Scientific Publishing.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2025. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*.

Artur Janicki and Krzysztof Szczypiorski. 2015. Why do older adults prefer some radio stations? helping to increase speech understanding. *J. Commun.*, 10(11):926–931.

Feng Jiang, Zhiyu Lin, Fan Bu, Yuhao Du, Benyou Wang, and Haizhou Li. 2025. S2s-arena, evaluating speech2speech protocols on instruction following with paralinguistic information. *arXiv preprint arXiv:2503.05085*.

Seunghee Kim, Ingyu Bang, Seokgyu Jang, Changhyeon Kim, Sanghwan Bae, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025a. Omhbench: Benchmarking balanced and grounded omni-modal multi-hop reasoning. *arXiv preprint arXiv:2508.16198*.

Seunghee Kim, Changhyeon Kim, and Taeuk Kim. 2025b. Fcmr: robust evaluation of financial cross-modal multi-hop reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23352–23380.

Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. 2023. Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Wentao Wang, Zhenghao Song, Dingling Zhang, and 1 others. 2025a. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*.

Chin-Jou Li, Kalvin Chang, Shikhar Bharadwaj, Eunjung Yeo, Kwanghee Choi, Jian Zhu, David Mortensen, and Shinji Watanabe. 2025b. Powsm: A phonetic open whisper-style speech foundation model. *arXiv preprint arXiv:2510.24992*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, and 1 others. 2024a. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, and 1 others. 2024b. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*.

Yunxin Li, Xinyu Chen, Shenyuan Jiang, Haoyuan Shi, Zhenyu Liu, Xuanyu Zhang, Nanhao Deng, Zhenran Xu, Yicheng Ma, Meishan Zhang, Baotian Hu, and Min Zhang. 2025c. Uni-moe-2.0-omni: Scaling language-centric omnimodal large model with advanced moe, training and data. *Preprint*, arXiv:2511.12609.

Heyang Liu, Yuhao Wang, Ziyang Cheng, Hongcheng Liu, Yiqi Li, Yixuan Hou, Ronghua Wu, Qunshan Gu, Yanfeng Wang, and Yu Wang. 2025. Vocalbench: Benchmarking the vocal conversational abilities for speech interaction models. *arXiv preprint arXiv:2505.15727*.

Timothy Liu and 1 others. 2022. Towards better characterization of paraphrases. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8592–8601.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760.

Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, and 1 others. 2025. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. *arXiv preprint arXiv:2512.02231*.

Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, and 1 others. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*.

OpenAI. 2025a. Gpt-5 system card. https://openai.com/index/gpt-5-system-card/. Accessed: 2026-03-11.

OpenAI. 2025b. Gpt-image-1. https://platform.openai.com/docs/models/gpt-image-1. Accessed: 2026-03-11.

OpenBMB. 2026. Minicpm-o series. https://github.com/OpenBMB/MiniCPM-o. Accessed: 2026-03-11.

Miles Purvis. 2025. English accent classifier (6 classes).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

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

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. The edinburgh international accents of english corpus: Towards the democratization of english asr. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, and 1 others. 2025. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue. *arXiv preprint arXiv:2510.13747*.

Chengyao Wang, Zhisheng Zhong, Bohao Peng, Senqiao Yang, Yuqi Liu, Haokun Gui, Bin Xia, Jingyao Li, Bei Yu, and Jiaya Jia. 2025a. Mgm-omni: Scaling omni llms to personalized long-horizon speech. *arXiv preprint arXiv:2509.25131*.

Ke Wang, Houxing Ren, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2025b. Voiceassistant-eval: Benchmarking ai assistants across listening, speaking, and viewing. *arXiv preprint arXiv:2509.22651*.

Wenbin Wang, Yang Song, and Sanjay Jha. 2024. Globe: A high-quality english corpus with global accents for zero-shot speaker adaptive text-to-speech. *arXiv preprint arXiv:2406.14875*.

Wiktionary contributors. 2026. Wiktionary, the free dictionary. https://en.wiktionary.org/wiki/Wiktionary:Main_Page. Accessed: 2026-03-11.

Arthur Wingfield and Julie L Ducharme. 1999. Effects of age and passage difficulty on listening-rate preferences for time-altered speech. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 54(3):P199–P202.

xAI. 2025. Grok 4. https://x.ai/news/grok-4. Accessed: 2026-03-11.

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

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

Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. 2025. Uro-bench: Towards comprehensive evaluation for end-to-end spoken dialogue models. *arXiv preprint arXiv:2502.17810*.

Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, and 1 others. 2025a. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10748–10757.

Shu-wen Yang, Ming Tu, Andy T Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. 2025b. Paras2s: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. *arXiv preprint arXiv:2511.08723*.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *Preprint*, arXiv:1904.01130.

Yue Zhao and Wei Lin. 2016. Study of the formant and duration in chinese whispered vowel speech. *Applied Acoustics*, 114:240–243.

Ziwei Zhou, Rui Wang, and Zuxuan Wu. 2025. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*.

Juan Pablo Zuluaga. 2023. Accent identification from speech recordings with ecapa-tdnn embeddings on commonaccent.

Statistics	Number
Total Instances	3,559
Emotion	597
angry	100
disgust	98
fear	100
joy	99
sadness	100
surprised	100
Global Accent	549
Australia	200
India	198
UK	151
Speech Rate	596
fast	298
slow	298
Timbre	596
adult female	147
adult male	150
elderly female	150
elderly male	149
Phonation	590
whisper	590
Pronunciation	631
heteronym	631
Text Generation LLM variants	3
Speech Paraphrasing LLM variants	3
Image Prompt Generation LLM variants	3
Image Generation Model variants	4
TTS Model variants	3
Avg. Image Resolution	512x512
Avg. Speech Duration (s)	6.8
Avg. Text Length (chars)	36.8

Table 5: Detailed statistics of OmniACBench.

Aspect	Metric	Value
Paraphrase	(MRPC) WPD↑ / LD↑	0.12 / 0.42
	(PAWS) WPD↑ / LD↑	0.07 / 0.13
	(Ours) WPD↑ / LD↑	0.11 / 0.69
Speech	WER↓ / CER↓	0.004 / 0.001
	STOI↑	0.994
Image	(keyword) CLIP↑ / LPIPS↑	0.067 / 0.373
	(Ours) CLIP↑ / LPIPS↑	0.124 / 0.466
Variants	Text Gen LLMs	3
	Paraphrase LLMs	3
	Image Prompt Gen LLMs	3
	Image Gen Models	4
	TTS models	3

Table 6: Quantitative verification of dataset construction quality. Higher WPD and LD denote greater paraphrase diversity; lower WER and CER and higher STOI indicate better speech quality; higher CLIP distance and LPIPS reflect greater image diversity. The Variants row reports the number of generation models used at each stage.

A Benchmark Construction Details

For **Speech Rate**, we define two target values, *fast* and *slow*. Image keywords for *fast* are built around situations that naturally imply urgency or emergency, such as *a brick falling toward a construction worker* or *a boiling-over pot*. In contrast, image keywords for *slow* are designed around elderly-directed interactions, such as scenes involving an *elderly person*. This was determined based on prior studies suggesting that speech directed to older adults should be delivered at a slower rate (Wingfield and Ducharme, 1999; Janicki and Szczypiorski, 2015; Chen et al., 2022b).

For **Pronunciation**, we focus on English heteronyms, where the same orthographic form has different pronunciations depending on meaning and context. The reference pronunciations were determined based on the entries provided in Wiktionary (Wiktionary contributors, 2026).

For **Phonation**, we focus on the target value *whisper*. The corresponding image keywords describe situations in which whispering is pragmatically appropriate, including *people studying in a library reading room* or *a quiet art museum with visitors walking slowly*. These scenes provide natural contextual grounding for suppressed vocal intensity and reduced voicing, making whisper-like phonation visually inferable.

For **Emotion**, we use six target values: *joy*, *surprised*, *angry*, *disgust*, *fear*, and *sadness*. The corresponding image keywords are centered on visually salient facial affective cues, such as *a person with an angry expression* or *a person with a disgusted expression*.

For **Global Accent**, we consider three target values: *India*, *UK*, and *Australia*. We construct image keywords not only from national symbols such as flags but also from culturally and geographically distinctive visual cues associated with each country. For example, keywords include *the national flag of India* or *the Opera House in Sydney, Australia*. This strategy increases visual diversity and encourages models to infer the relevant national context from broader cultural grounding rather than from a single canonical cue.

For **Timbre**, we define four target values: *adult male*, *adult female*, *elderly male*, and *elderly female*. Image keywords are constructed to foreground visually recognizable cues of speaker age and gender, such as *an elderly woman*.

Stage	Models
Text Generation	Gemini 3 Flash, GPT 5 Nano, Claude Haiku 4.5
Paraphrasing	Gemini 3 Flash, GPT 5 Nano, Claude Haiku 4.5
Image Prompt Generation	Gemini 3 Flash, GPT 5 Nano, Claude Haiku 4.5
Image Generation	Gemini 3 Pro Image, Gemini 2.5 Flash Image, GPT Image 1, GPT Image 1.5
Text-to-Speech	Eleven Multilingual v2, Eleven Flash v2.5, Eleven Turbo v2.5
LLM-based Filtering	Grok 4 Fast

Table 7: Model pools used in dataset construction. For each generation call, one model is randomly sampled from the corresponding stage-specific pool. LLM-based filtering is performed using a separate model.

B Quality Control Protocol Details

B.1 Data Filtering

To ensure that OmniACBench instances faithfully reflect the intended benchmark design, we apply three filtering criteria at different stages of the construction pipeline: **Semantic Preservation**, **Text Neutrality**, and **Image–Keyword Alignment**. These criteria are designed to remove artifacts that could otherwise undermine the validity of the evaluation, such as semantically drifted instructions, scripts that leak the target acoustic value, or images that no longer reflect their source concepts. For each criterion, we first conduct LLM-based filtering as a scalable first-pass screening step, and then perform human verification using the same criterion to confirm the final decision.

Semantic Preservation is applied to spoken instructions generated through paraphrasing. Its purpose is to ensure that each paraphrased instruction remains faithful to the meaning and intent of the original template, without introducing semantic drift, unintended constraints, or changes in the target acoustic dimension. This step is important because the spoken instruction serves as the control signal for the task; if its meaning changes during paraphrasing, the resulting instance may no longer represent the intended evaluation condition. The prompt used for this filtering step is shown in Figure 8.

Text Neutrality is applied to generated scripts in order to prevent shortcut solutions. Specifically, we remove scripts whose lexical content alone reveals the intended target value, since such cases would allow models to recover the desired acoustic realization directly from the text rather than from multimodal context. This filtering criterion helps preserve the intended role assignment of the benchmark, where the text provides only the verbal content to be spoken and does not itself encode the acoustic target. The prompt used for this filtering step is shown in Figure 9.

Image–Keyword Alignment is applied to generated images to verify that they remain semantically consistent with the original image keywords used during construction. Since images are synthesized from expanded prompts rather than directly from the initial keywords, this step checks whether the final image still preserves the intended scene or concept associated with the target acoustic value. This is necessary to ensure that the visual modality provides valid contextual grounding, rather than introducing irrelevant or misleading content. The prompt used for this filtering step is shown in Figure 10.

B.2 Quantitative Verification

Paraphrasing Quality Because spoken instructions in OmniACBench are generated by paraphrasing feature-level templates, it is important to verify that they exhibit sufficient linguistic diversity rather than remaining close to a small set of fixed phrasings. Following prior benchmark construction work (Kim et al., 2025b,a), we measure this property using Word Position Deviation (WPD) and Lexical Deviation (LD) (Liu et al., 2022). OmniACBench achieves WPD/LD scores of 0.11/0.69, compared to 0.12/0.42 on MRPC (Dolan and Brockett, 2005) and 0.07/0.13 on PAWS (Zhang et al., 2019). These results indicate that the spoken instructions in OmniACBench preserve substantial lexical diversity while avoiding overly rigid template repetition.

Speech Quality We also verify the quality of synthesized speech, since unintelligible or text-inaccurate audio would introduce noise unrelated to the intended benchmark capability. To this end, we evaluate transcription fidelity using WER and CER computed with Whisper-large-v3 (Radford et al., 2022), and intelligibility using STOI (↑) (Kumar et al., 2023). The resulting scores are WER 0.004, CER 0.001, and STOI 0.994, indicating that the synthesized speech remains highly faithful to the intended script while also being near-perfect in

intelligibility.

Diversity Control We first evaluate whether meta-prompting improves image diversity over naïve keyword prompting. For each keyword group, we compare images generated by the two approaches using average pairwise CLIP embedding cosine distance (Radford et al., 2021) and LPIPS (Zhang et al., 2018), which capture semantic and perceptual variation, respectively. The meta-prompt approach yields higher diversity on both metrics (CLIP 0.1242 vs. 0.0671; LPIPS 0.4661 vs. 0.3733), showing that richer prompt expansion leads to greater visual variation than keyword-only prompting. A qualitative comparison of images generated by the two approaches for the same keyword is provided in Figure 17.

Second, to reduce model-specific bias and increase generation diversity, we maintain model pools for each stage of tri-modal generation and randomly sample one model per instance from the corresponding pool. For text generation, paraphrasing, and image prompt generation, we employ three LLMs: Gemini 3 Flash (Google, 2025b), GPT 5 Nano (OpenAI, 2025a), and Claude Haiku 4.5 (Anthropic, 2025). For image generation, we utilize four models: Gemini 3 Pro Image (Google, 2025c), Gemini 2.5 Flash Image (Google, 2025a), GPT Image 1, and GPT Image 1.5 (OpenAI, 2025b). For TTS, we use three models: Eleven Multilingual v2 (ElevenLabs, 2026b), Eleven Flash v2.5 (ElevenLabs, 2026a), and Eleven Turbo v2.5 (ElevenLabs, 2026c). At each generation stage, one model is randomly selected from the corresponding pool to increase diversity in the generated text, images, and speech. An independent model, Grok 4 Fast (xAI, 2025), is used for LLM-based filtering.

C Evaluator Training Details

Rather than relying on off-the-shelf evaluators, we train task-specific classifiers for all three abstract features. This choice is motivated by both label-space mismatch and empirical performance gaps. On our held-out test sets for Emotion, Global Accent, and Timbre, off-the-shelf alternatives do not consistently align with the target categories or provide sufficient accuracy: emotion2vec (Ma et al., 2024) achieves 84.57% on Emotion, while ECAPA-TDNN-based English accent classifier (Zuluaga, 2023) and Wav2Vec2 based English accent classifier (Purvis, 2025) achieve 51.83% and 59.91%, respectively, on Global Accent, all below our task-specific evaluators. For Timbre, moreover, we are not aware of a readily available off-the-shelf evaluator that directly matches our four-way setting (*adult/elderly* $\times$ *female/male*). Task-specific training therefore provides a better label match and a more reproducible evaluation pipeline for OmniACBench.

Data Collection We train task-specific classifiers for three abstract acoustic features: Emotion, Global Accent, and Timbre. For **Emotion**, we aggregate speech data from four publicly available datasets—CREMA-D (Cao et al., 2014) (6,355 samples), TESS (Dupuis and Kathleen Pichora-Fuller, 2011) (2,400), RAVDESS (Livingstone and Russo, 2018) (1,888), and SAVEE (Haq and Jackson, 2011) (360)—yielding 11,003 samples in total across six classes (*angry, disgust, fear, joy, sadness, surprised*). For **Global Accent**, we collect 12,000 samples from GLOBE (Wang et al., 2024) (7,334), EDACC (Sanabria et al., 2023) (2,004), Common-Accent (DTU54DL) (1,497), and MINDS-14 (Gerz et al., 2021) (1,165), covering three classes (*Australian, Indian, UK*). For **Timbre**, we collect 12,000 samples from GLOBE (Wang et al., 2024) (11,400) and EDACC (Sanabria et al., 2023) (600), covering four classes (*adult female, adult male, elderly female, elderly male*), where *adult* refers to speakers in their 20s–30s and *elderly* refers to speakers aged 60 and above. All three datasets are split into train/dev/test partitions.

Training All classifiers use WavLM-Large (Chen et al., 2022a) as the backbone, with a task-specific classification head on top of mean-pooled frame representations. We fully fine-tune the backbone using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 10^{-4}), a cosine learning-rate schedule with 2-epoch linear warmup, label smoothing of 0.1, and gradient clipping at norm 1.0. Training runs for 30 epochs with a per-GPU batch size of 32. We select the final configuration via grid search over head learning rate $\in \{10^{-5}, 10^{-4}, 10^{-3}\}$ and backbone learning rate $\in \{5 \times 10^{-7}, 5 \times 10^{-6}, 5 \times 10^{-5}\}$, choosing the checkpoint with the best development-set accuracy. The selected learning rates are 10^{-4} for the head and 5×10^{-5} for the backbone for all three classifiers. The resulting checkpoints achieve accuracies of 89.43%, 97.29%, and 96.67% on the held-out test sets for Emotion, Global Accent, and Timbre, respectively, supporting their use as automatic evaluators in OmniACBench.

D Context Flow Analysis Details

For the context-flow analysis in Section 5.3, we compared MiniCPM-o 4.5 (OpenBMB, 2026) and Qwen3-Omni 30B (Xu et al., 2025b) to examine whether information about the intended acoustic attribute remains decodable as it propagates through each model toward speech generation. In MiniCPM-o 4.5, representations were extracted from the LLM backbone, the intermediate projection stage, and the TTS decoder; in Qwen3-Omni 30B, they were extracted from the Thinker, the intermediate projection stages, and the Talker. We then trained a separate linear probe for each layer and each acoustic feature—Emotion, Global Accent, and Timbre—to predict the target subcategory from the corresponding hidden representation. Each probe was implemented as a single linear classifier trained with cross-entropy loss and AdamW. Evaluation was conducted using repeated stratified train/test splits, with hidden representations standardized using training-set statistics only. Performance was measured using balanced accuracy (i.e., mean per-class recall) on the held-out split, averaged across repeated runs.

E Model Details

E.1 Open-source Omni Models

We evaluate eight omni models that support text, image, and speech inputs and generate speech outputs. This section focuses on architecture-level characteristics relevant to context-grounded acoustic control.

MiniCPM-o 4.5 (OpenBMB, 2026). The MiniCPM-o family adopts an end-to-end omni design that directly connects modality encoders and decoders with a language backbone through hidden-state interactions, rather than a cascaded ASR→LLM→TTS stack. Its architecture combines SigLip2 for vision, Whisper-medium for audio understanding, CosyVoice2-style speech tokenization and Token2Wav, and a Qwen3-8B backbone, together with full-duplex streaming and timeline modeling (TDM) for real-time interaction. For speech specifically, input audio is encoded online and output speech is generated by an interleaved text–speech token decoder, enabling synchronized full-duplex response generation.

InteractiveOmni (Tong et al., 2025). InteractiveOmni integrates a vision encoder, an audio encoder, a language model, and a speech decoder into a unified architecture for audio-visual multi-turn dialogue. A central design element is multi-stage training with dedicated data curation for long-horizon conversational context and speech-oriented response quality. Its speech pathway is explicitly end-to-end: audio input is handled by the unified audio encoder, while speech output is produced by the integrated speech decoder rather than by an external TTS post-processor.

Qwen3-Omni 30B (Xu et al., 2025b). Qwen3-Omni 30B uses a Thinker–Talker Mixture-of-Experts architecture that separates high-level reasoning from speech generation. For speech input, it adopts AuT, a lightweight continuous audio encoder with a 12.5Hz frame rate that maps audio into semantic features for the Thinker. For speech output, the Talker autoregressively predicts multi-codebook discrete speech codec tokens and reconstructs waveform with a lightweight causal ConvNet (Code2Wav), targeting low first-packet latency and streaming generation.

Qwen2.5-Omni (Xu et al., 2025a). Qwen2.5-Omni introduces an end-to-end omni stack with block-wise audio and video streaming encoders and temporal multimodal positional encoding (TM-RoPE) for synchronized time modeling. It also adopts a Thinker–Talker paradigm: speech input is consumed through the block-wise streaming audio encoder, while speech output is produced as discrete speech tokens via a dual-track autoregressive Talker and then rendered through a sliding-window DiT-based codec pathway for streaming audio.

Uni-MoE-2.0-Omni (Li et al., 2025c). Uni-MoE-2.0-Omni emphasizes scalable multimodal routing via a dynamic-capacity MoE design with shared, routed, and null experts, coupled with omni-modality 3D-RoPE for unified spatiotemporal representation. Its training pipeline includes progressive stages from omni pretraining to preference optimization. For speech handling, the architecture includes a unified speech encoder for audio feature extraction and dedicated speech-generation tokens that synchronize speech and text generation, combined with a context-aware MoE-TTS stage for high-quality synthesis.

MGM-Omni 7B (Wang et al., 2025a). MGM-Omni-7B proposes a brain–mouth dual-track token architecture that decouples reasoning from low-latency speech generation. In this design, speech

input is processed by a dual audio encoder for robust understanding, while speech output is generated by chunk-based parallel decoding to narrow the text/speech token-rate gap and improve realtime responsiveness.

E.2 Commercial Models

For diversity control in Section 3.3.2, we use commercial model pools at each generation stage. Since full implementation details are often undisclosed for closed models, we summarize publicly documented architectural signals and distinctive characteristics.

LLMs for text generation, paraphrasing, and meta-prompting. Gemini 3 Flash (Google, 2025b) is a Gemini 3 family model with configurable thinking budgets and long-context multimodal input. GPT 5 Nano (OpenAI, 2025a) belongs to the GPT 5 family with explicit fast/thinking model routing and a small reasoning variant (`gpt-5-thinking-nano`). Claude Haiku 4.5 (Anthropic, 2025) is positioned as a low-latency model in the Claude 4.5 line with extended-thinking mode and large-context handling.

Image generation models. Gemini 3 Pro Image (Google, 2025c) and Gemini 2.5 Flash Image (Google, 2025a) provide native image generation and editing within Gemini multimodal models; Gemini 2.5 Flash additionally adopts hybrid reasoning and sparse-MoE-based scaling. GPT Image 1 and GPT Image 1.5 (OpenAI, 2025b) are OpenAI's natively multimodal image-generation models supporting text-and-image conditioning and iterative editing workflows.

Text-to-speech models. Eleven Multilingual v2 (ElevenLabs, 2026b) focuses on expressive multilingual speech quality. Eleven Flash v2.5 (ElevenLabs, 2026a) targets very low latency for realtime use, while Eleven Turbo v2.5 (ElevenLabs, 2026c) balances speed and quality for general-purpose synthesis.

LLM-based filtering model. We use Grok 4 Fast (xAI, 2025) as a separate filtering model. The Grok 4 family is characterized by large-scale reinforcement learning and native tool-use integration, which makes it suitable for independent quality filtering without sharing generation-stage model biases.

F Experimental Environment

All experiments were conducted on a machine equipped with Intel Xeon Gold 6338 CPU @ 2.00GHz (2 sockets $\times$ 32 cores, up to 3.20GHz boost), and 4 $\times$ NVIDIA A100-SXM4 GPUs each with 80 GB of memory. The system ran Ubuntu 20.04.5 LTS with CUDA compilation tools release 11.8. During both dataset generation and evaluation, the random seed was fixed to 42 to ensure reproducibility.

Text Generation Prompt

Your task is to generate a single line of dialogue.

1. The line must be emotionally neutral.
2. It must be nationally neutral.
3. It must be neutral with respect to gender and age.
4. The line must be a declarative sentence (not a question or exclamation).

Output only the single line of dialogue with no additional text.

Figure 5: A prompt used for text transcript generation.

Instruction Paraphrasing Prompt

Paraphrase the provided instruction template.

Condition 1: The original and the paraphrased instruction must have exactly the same essential meaning.
Condition 2: Perform paraphrasing with consideration of lexical variation.
Condition 3: Perform paraphrasing with consideration of syntactic variation.

Output only the single paraphrased instruction with no additional text.

original instruction template: {original_template}

paraphrased instruction:

Figure 6: A prompt used for instruction paraphrasing.

Image Meta-Prompt

You are a professional prompt engineer who writes prompts for image generation models.

Create a single image generation prompt that satisfies all of the following conditions:

1. The element **{image_keyword}** must be the central focus of the image and be strongly emphasized and clearly visible in the prompt.
2. The final output must be only one image generation prompt and nothing else.
3. All elements must be separated by commas.
4. The prompt must consist of 5 to 8 comma-separated elements that are visually clear and specific.

Generate the image generation prompt that meets these conditions.

Figure 7: A meta-prompt template used for image generation prompt expansion.

Semantic Preservation Filtering Prompt

original instruction: {original}
paraphrased instruction: {paraphrased}
If the original instruction and the paraphrased instruction have the same meaning, output Yes; otherwise, output No.
Output only Yes or No without any additional explanation.
Answer:

Figure 8: A prompt used for LLM-based semantic preservation filtering.

Text Neutrality Filtering Prompt

transcript: {transcript}
Target Acoustic Feature: {Target_Acoustic_Feature}
Can the Target Acoustic Feature be fully inferred using only the provided transcript? Answer Yes or No.
Output only Yes or No without any additional explanation.
Answer:

Figure 9: A prompt used for LLM-based text neutrality filtering.

Image-Keyword Alignment Filtering Prompt

Does the provided image depict "{image_keyword}"? Answer Yes or No.
Output only Yes or No without any additional explanation.
Answer:

Figure 10: A prompt used for LLM-based image-keyword alignment filtering.

Human Verification Instruction: Semantic Preservation

original instruction: {original}
paraphrased instruction: {paraphrased}
Do the original instruction and the paraphrased instruction have the same meaning? Answer Yes or No.

Figure 11: An instruction used for human-based semantic preservation filtering.

Human Verification Instruction: Text Neutrality

transcript: {transcript}
Target Acoustic Feature: {Target_Acoustic_Feature}
Can the Target Acoustic Feature be fully inferred using only the provided transcript? Answer Yes or No.

Figure 12: An instruction used for human-based text neutrality filtering.

Human Verification Instruction: Image-Keyword Alignment

Does the provided image depict "{image_keyword}"? Answer Yes or No.

Figure 13: An instruction used for human-based image-keyword alignment filtering.

Human Annotator Instruction: Emotion

Speech: {speech}
Listen to the following speech and identify the emotion expressed by the speaker.
(1) Angry (2) Disgust (3) Fear (4) Joy (5) Sadness (6) Surprised

Human Annotator Instruction: Global Accent

Speech: {speech}
Listen to the following speech and identify the English accent of the speaker.
(1) India (2) UK (3) Australia

Human Annotator Instruction: Timbre

Speech: {speech}
Listen to the following speech and identify the gender and age group of the speaker.
(1) Adult Female (2) Adult Male (3) Elderly Female (4) Elderly Male

Figure 14: Human annotator instructions for abstract acoustic feature validation used in Section 5.2.

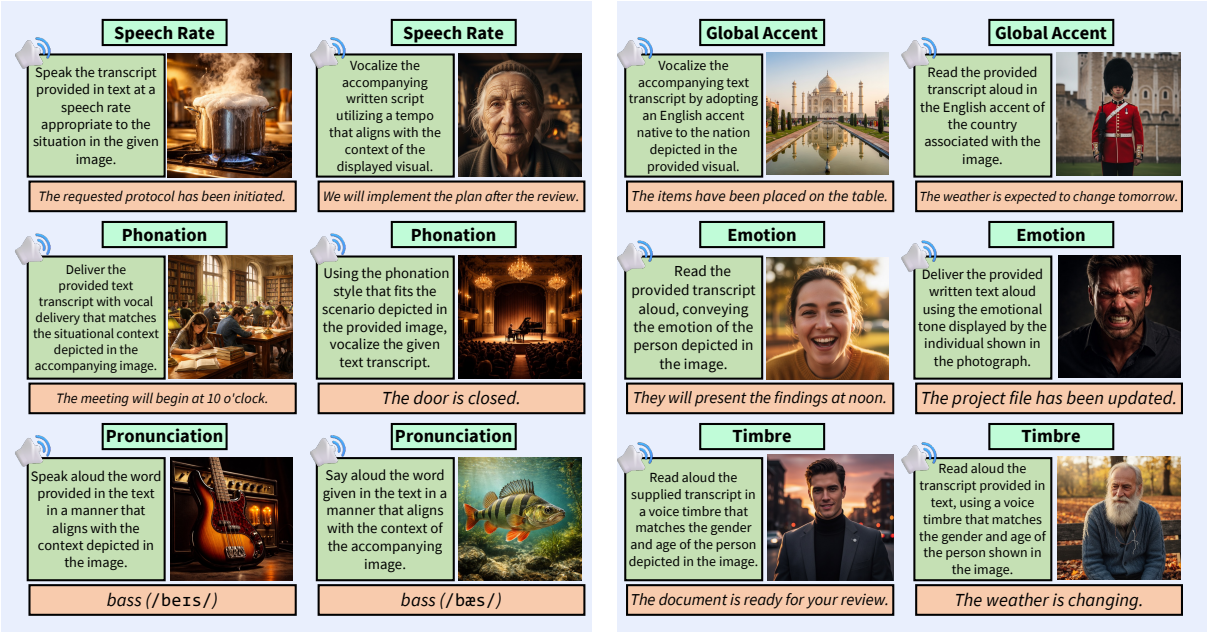

Figure 15: Examples of OmniACBench data for each acoustic feature.

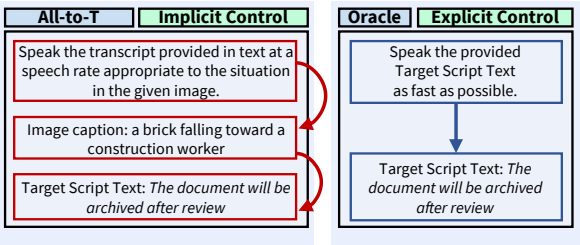

Figure 16: Illustration of the All-to-T and Oracle conditions used in Controlled Input Decomposition. All-to-T textualizes all inputs while still requiring inference of the target acoustic value from context, whereas Oracle makes the target value explicit in the instruction.

Figure 17: Image diversity comparison between keyword-based and meta-prompt generation for the keyword "a person with a sad expression." All images in this example are generated with Gemini 2.5 Flash Image.

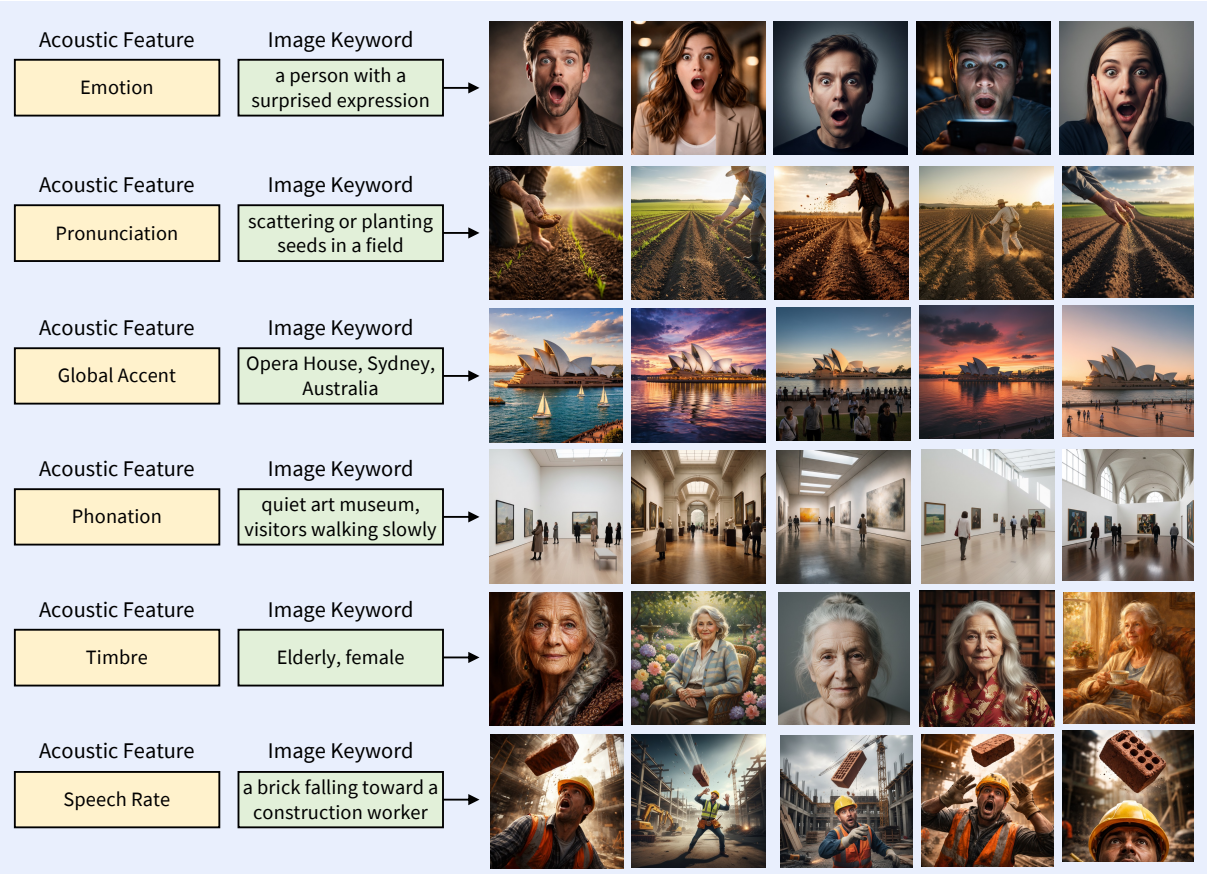

Figure 18: Examples of image keywords and representative final generated images for each image keyword.

Acoustic Feature	Instruction Template	Paraphrased Instruction
Emotion	Speak the transcript provided in text with the emotion of the person in the given image.	Read the provided transcript aloud, adopting the emotional tone of the person shown in the image.
Global Accent	Speak the transcript provided in text using the English accent of the country related to the given image.	Read aloud the provided transcript in the English accent of the country associated with the image.
Phonation	Speak the transcript provided in text using a phonation style appropriate to the situation in the given image.	Vocalize the textual content supplied by employing a voice delivery method that matches the contextual setting depicted in the provided picture.
Pronunciation	Pronounce the word provided in text in a way that matches the context of the given image.	Pronounce aloud the word supplied in the text in a way that aligns with the context of the accompanying image.
Timbre	Speak the transcript provided in text with a timbre appropriate to the gender and age of the person in the given image.	Deliver the provided text transcript using a voice tone that matches the gender and age characteristics visible in the supplied image.
Speech Rate - fast	Speak the transcript provided in text at a speech rate appropriate to the situation in the given image.	Vocalize the supplied text script, adjusting your pacing to match the context depicted in the accompanying visual.
Speech Rate - slow	Speak the transcript provided in text at a speech rate appropriate to the subject in the given image.	Read the supplied text transcript aloud at a pace appropriate to the subject depicted in the image.

Figure 19: Example paraphrases of spoken instruction templates.

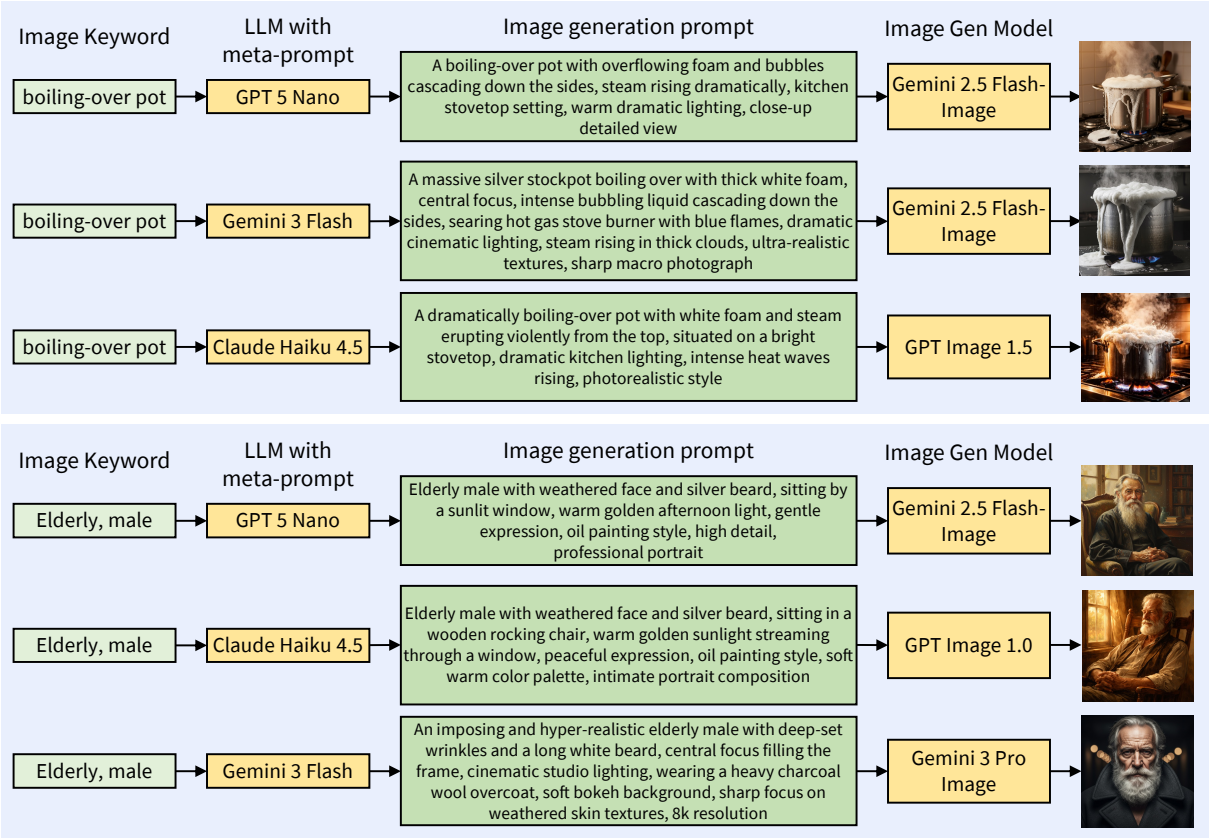

Figure 20: Examples of meta-prompt-based image prompt expansion.

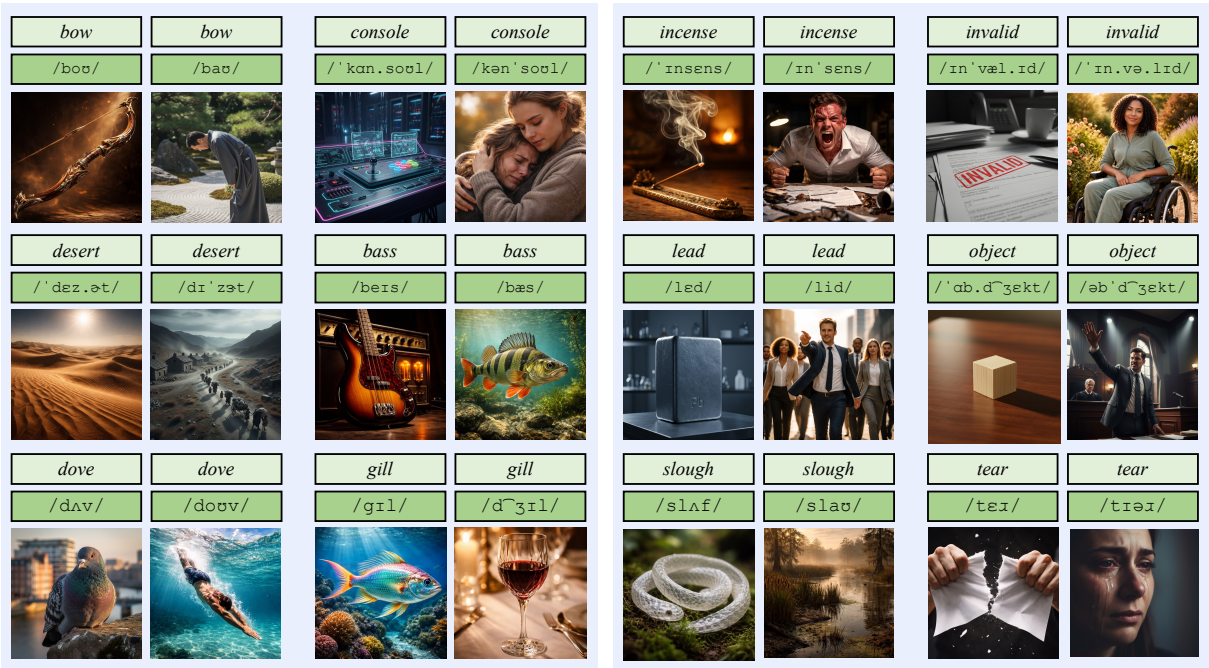

Figure 21: Image examples of pronunciation features.