Beyond Transcription: Unified Audio Schema for Perception-Aware AudioLLMs

Anonymous ACL submission

Abstract

Recent Audio Large Language Models (AudioLLMs) exhibit a striking performance inversion: while excelling at complex reasoning tasks, they consistently underperform on fine-grained acoustic perception. We attribute this gap to a fundamental limitation of ASR-centric training, which provides precise linguistic targets but implicitly teaches models to suppress paralinguistic cues and acoustic events as noise. To address this, we propose Unified Audio Schema (UAS), a holistic and structured supervision framework that organizes audio information into three explicit components—Transcription, Paralinguistics, and Non-linguistic Events—within a unified JSON format. This design achieves comprehensive acoustic coverage without sacrificing the tight audio-text alignment that enables reasoning. We validate the effectiveness of this supervision strategy by applying it to both discrete and continuous AudioLLM architectures. Extensive experiments on MMSU, MMAR, and MMAU demonstrate that UAS-Audio yields consistent improvements, boosting fine-grained perception by 10.9% on MMSU over same-size state-of-the-art models while preserving robust reasoning capabilities.

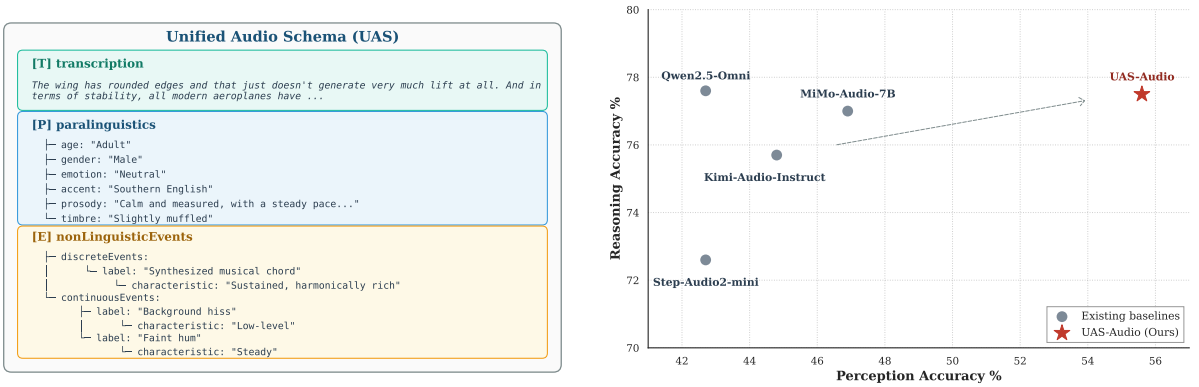
1 Introduction

Recent Audio Large Language Models (AudioLLMs) present a striking paradox: models capable of complex reasoning often fail at elementary auditory perception. While they excel on reasoning-heavy benchmarks, they struggle to reliably identify speaker traits, emotion, prosody, or even simple non-linguistic acoustic events (Yang et al., 2024; Wang et al., 2025; Kwon et al., 2025). For instance, on the MMSU benchmark, current AudioLLMs achieve approximately 70% accuracy on complex reasoning tasks yet drop sharply to around 40% on fundamental perception tasks. Such perceptual blind spots lead to practical failures: a model may correctly transcribe "I'm fine" while completely missing the trembling voice that signals distress, or fail to notice a door slam that signals an abrupt end to the interaction.

Despite the rapid scaling of both language backbones and audio encoders, these perceptual failures persist across different model sizes and architectures. This persistence suggests that the underlying issue is unlikely to be solely a result of insufficient model capacity or architectural limitations. Instead, it points to a more systemic factor shared by most existing AudioLLMs: how audio information is supervised during training.

Most existing models rely heavily on automatic speech recognition (ASR) as their training signal, using text transcription as the primary interface between audio and language. While effective for semantic alignment, ASR is inherently selective: to recover canonical text, it deliberately normalizes away prosody, speaker identity, emotion, and acoustic context. This training objective creates a fundamental asymmetry—models are consistently rewarded for reasoning about *what* is said, while being implicitly discouraged from attending to *how* it is said or what else occurs acoustically. As a result, perception is not merely under-trained — it is systematically de-emphasized.

Motivated by this perspective, we introduce the **Unified Audio Schema (UAS)**, a structured textual representation that decomposes audio into three complementary components (Figure 1a): *Transcription* captures the spoken content; *Paralinguistics* encodes speaker-level attributes such as emotion, age, gender, accent, prosody, and timbre; and *Non-linguistic Events* describes the acoustic context, including both discrete sounds (e.g., door slams, laughter) and continuous background conditions (e.g., ambient noise, music). This decomposition follows the classical taxonomy of speech information (Laver, 1994) and is designed to expose perceptual dimensions explicitly rather than



(a) Unified Audio Schema (UAS)

(b) Performance of UAS-Audio

Figure 1: **Overview of the Unified Audio Schema (UAS) and evaluation results.** (a) UAS structures audio information into three components: Transcription, Paralinguistics, and Non-linguistic Events. (b) Reasoning vs. Perception accuracy on MMSU. UAS-Audio significantly enhances perception while maintaining robust reasoning.

implicitly. Leveraging this structured format, UAS enables AudioLLMs to retain perceptual information without sacrificing semantic alignment.

Crucially, this schema does not require expensive manual annotation. We demonstrate that UAS training data can be automatically synthesized at scale from existing corpora using off-the-shelf models, converting standard ASR datasets into rich, perception-aware supervision.

Leveraging this scalable pipeline, we validate the effectiveness of UAS on both continuous (Xu et al., 2025a; Wu et al., 2025) and discrete (Zeng et al., 2024) architectures. Experiments confirm that restructuring supervision yields consistent gains across both paradigms. As shown in Figure 1b, UAS-Audio achieves an 11% absolute improvement in perception accuracy over state-of-the-art baselines on MMSU, while strictly preserving reasoning capabilities. Beyond perception, UAS-Audio demonstrates robust generalization across diverse audio domains, achieving state-of-the-art performance on the MMAR reasoning benchmark (60.1%) and competitive results on MMAU, securing the highest average score (65.2%) across all three benchmarks.

In summary, our contributions are threefold:

- We identify that the perceptual weakness of current AudioLLMs stems from ASR-centric supervision, which systematically deemphasizes paralinguistic and non-linguistic information during training.
- We propose the Unified Audio Schema (UAS), a structured representation that explicitly decomposes audio into transcription, paralinguistics, and non-linguistic events, along with a scalable pipeline to synthesize UAS annotations using off-the-shelf expert models.
- We train UAS-Audio and demonstrate consistent improvements across both continuous and discrete AudioLLM architectures, achieving an absolute 11% gain in perception accuracy on MMSU benchmark while preserving reasoning performance.

2 Unified Audio Schema (UAS)

2.1 Schema Definition and Rationale

Our schema design is grounded in the semiotic framework of speech signals defined by Laver (1994). According to Laver, speech is not a monolithic stream but a composite of three information layers: (1) the Linguistic Layer, conveying the abstract semantic message; (2) the Paralinguistic Layer, encompassing voluntary vocal variations (e.g., tone, prosody) that signal attitude; and (3) the Extralinguistic Layer, containing indexical features identifying biological constraints (e.g., age, gender) and the physical environment.

Building upon this hierarchy, we aim to preserve the full spectrum of acoustic nuance by explicitly disentangling the signal's semantic and non-semantic components. We map Laver's theoretical layers into a practical annotation schema organized along three functional dimensions:

Transcription. Corresponding to the linguistic layer, this field preserves the verbatim speech content with 100% linguistic fidelity equivalent to ASR

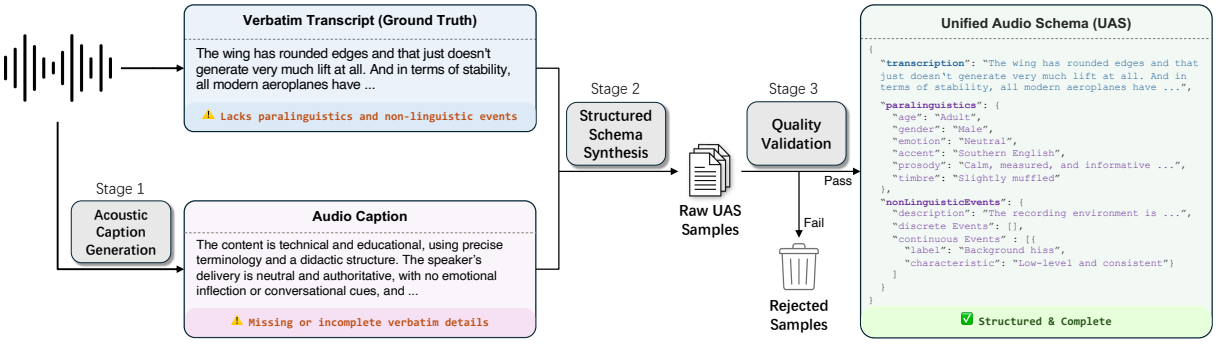


Figure 2: Overview of the UAS Data Generation Pipeline. (Stage 1) Acoustic captions are generated to capture paralinguistic and environmental context. (Stage 2) A structured synthesizer merges these captions with ground-truth transcripts. (Stage 3) An automated validation stage filters hallucinations and enforces logical consistency, yielding a final UAS that is 'Structured & Complete'

output. This field ensures that UAS never sacrifices the semantic precision compared to standard ASR.

Paralinguistics. Capturing the paralinguistic layer and the speaker-intrinsic aspects of the extralinguistic layer, this field consists of structured attributes describing *how* speech is produced. It is organized into six explicit subfields: (1) *Age*: Speaker age group (e.g., "Adult", "Child", "Elderly"); (2) *Gender*: Speaker gender (e.g., "Male", "Female"); (3) *Emotion*: Emotional state (e.g., "Neutral", "Happy", "Angry", "Sad"); (4) *Accent*: Regional or linguistic accent (e.g., "Southern American English"); (5) *Prosody*: Speaking style description including pace, intonation, and rhythm; and (6) *Timbre*: Voice quality characteristics describing the acoustic texture of the speaker's voice (e.g., "Slightly muffled", "Clear and resonant").

Non-linguistic Events. Representing the environmental aspect of the extralinguistic layer, this field gives acoustic information beyond speech, capturing the auditory scene through three subfields: (1) *Description*: Overall characterization of the recording environment; (2) *Discrete Events*: Identifiable sound occurrences with clear temporal boundaries (e.g., door slam); and (3) *Continuous Events*: Ambient sounds persisting throughout the audio (e.g., engine rumble). Each discrete or continuous event is annotated with a label and characteristic (e.g., "stylus click" with "sharp, high-pitched", "Background hiss" with "low-level and consistent").

It is important to note that UAS supports both speech and non-speech data (e.g., pure environment sounds or music). For audio segments containing no human speech, the "transcription" field and all relevant paralinguistic subfields are set to "null".

This structured decomposition offers critical advantages beyond mere organization. First, it facilitates **disentangled learning** by transforming the implicit task of "holistic understanding" into explicit subtasks. This prevents feature conflation, forcing the model to distinguish *what* is said from *how* it is said. Second, the schema introduces **syntactic invariance**. Unlike unstructured captions that suffer from high entropic variability (i.e., many ways to describe the same sound), UAS provides a consistent, low-entropy supervision target, thereby reducing learning difficulty and stabilizing optimization. Finally, it ensures **programmatic accessibility**. The rigorous JSON format bridges the gap between the probabilistic nature of LLMs and the deterministic requirements of software interfaces, allowing downstream applications to reliably utilize acoustic attributes without complex parsing.

2.2 Scalable UAS Data Generation Pipeline

To operationalize UAS at scale, we develop an automated pipeline that transforms existing ASR corpora into UAS-formatted supervision. The pipeline proceeds in three stages, as illustrated in Figure 2.

Stage 1: Acoustic Caption Generation. Given raw audio samples with their original transcriptions, we first employ a caption model (Xu et al., 2025b) as an acoustic captioner to generate rich descriptions of paralinguistic attributes and environmental sounds. The captioner is prompted to describe speaker characteristics (age, gender, emotion, accent, prosody, timbre), and non-speech acoustic events (e.g., background environment, discrete sounds, continuous ambient sounds or music) in detail. This stage extracts the acoustic information that ASR supervision inherently discards.

Stage 2: Structured Schema Synthesis. The generated acoustic captions are then combined with ground-truth transcriptions and processed through an LLM to synthesize structured UAS annotations. We design a carefully hand-crafted prompt (detailed in Appendix H) that instructs the LLM to: (1) preserve the original transcription verbatim in the designated field; (2) extract and normalize paralinguistic attributes from the caption into the predefined categories (age, gender, emotion, accent, prosody, timbre); and (3) organize non-linguistic event descriptions into the hierarchical schema with discrete and continuous event separation. This synthesis step ensures both semantic fidelity (via ground-truth transcription) and acoustic completeness (via caption-derived attributes).

Stage 3: Quality Validation. To ensure strict data reliability, we implement a multi-level automated validation pipeline: (1) Ontology Constraints Enforcement: All categorical fields (e.g., emotion, gender) are verified against a predefined closed-set vocabulary, discarding synonyms or open-ended hallucinations from the LLM. (2) Transcription Integrity: Exact string matching with the ground truth ensures zero semantic loss. (3) Logical Consistency Filtering: Rule-based checks resolve inter-field conflicts, such as strictly mapping empty transcriptions to "null" paralinguistic fields, and rejecting samples where gender attributes contradict acoustic timbre (e.g., "Male" label with "High-pitched feminine" description). (4) Duration-Content Alignment: Heuristic filters discard samples where the description's complexity disproportionately exceeds the audio duration, reducing the risk of generative hallucination.

Human Verification of Data Quality. To verify the automated pipeline, we manually audited $N = 200$ random samples across speech, music, and environmental sounds. The results demonstrate high reliability, with most paralinguistic and environmental attributes achieving a mean accuracy of over 95%. Detailed methodology and per-field accuracy scores are provided in Appendix A.

2.3 UAS-QA

Beyond raw UAS annotations, we additionally synthesize a **UAS-QA** dataset to train the model to leverage structured acoustic knowledge for downstream tasks. Based on UAS annotations, we automatically generate three types of question-answer pairs: (1) **Direct QA**: Questions querying specific UAS fields (e.g., "What is the speaker's emotion?" → "Neutral"; "What is the speaker's accent?" → "Southern American English"); (2) **Multiple Choice**: Questions with candidate options derived from the UAS attribute vocabulary (e.g., "What is the speaker's age group? A. Child B. Adult C. Elderly" → "B. Adult"); and (3) **Yes/No Questions**: Binary verification questions (e.g., "Is the speaker male?" → "Yes"; "Are there discrete sound events in this audio?" → "No").

This diverse question format ensures comprehensive coverage of all UAS fields—transcription, paralinguistics, and non-linguistic events—while providing explicit supervision signals that encourage the model to attend to fine-grained acoustic details during inference.

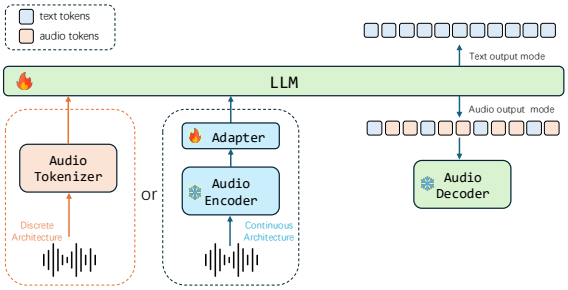


Figure 3: **Overview of UAS-Audio architectures**. Audio input is processed via *either* discrete tokenization *or* continuous encoding (left). The LLM generates text-only or bi-modal outputs (right).

3 UAS-Audio

3.1 Overview

To validate the effectiveness of UAS supervision, we develop **UAS-Audio** (Figure 3), an AudioLLM designed to address the perception–reasoning imbalance identified in existing models. In this section, we focus on the continuous architecture of UAS-Audio, whose design aims to preserve the strong reasoning capabilities of current AudioLLMs while substantially enhancing fine-grained acoustic perception—achieving holistic audio understanding without the trade-offs inherent in ASR-centric or caption-based training paradigms. The discrete variant (denoted as UAS-Audio-D), which follows a similar design philosophy with architecture-specific adaptations, is detailed in Appendix B.

Following the successful paradigm established by recent AudioLLMs (Xu et al., 2025a; Wu et al., 2025), the continuous UAS-Audio model adopts a

four-component framework that has proven effective for audio–language modeling: (1) an **Audio Encoder** that transforms raw waveforms into continuous representations; (2) a **Projection Layer** that aligns audio representations with the language model's embedding space; (3) a **Large Language Model** backbone that performs reasoning over combined audio–text inputs; and (4) a **Speech Decoder** based on the flow matching architecture (Lipman et al., 2023), which converts audio tokens into mel-spectrograms that are subsequently transformed into waveforms using a HiFi-GAN vocoder (Kong et al., 2020).

We next describe the training pipeline for this continuous architecture.

3.2 Training Pipeline

We follow the standard multi-stage alignment protocol established in Wu et al. (2025); Xu et al. (2025a). We do not introduce new modules or specialized loss functions; we simply plug in the UAS data.

Stage 1: Discrete Token Alignment. Although our primary architecture leverages a continuous audio encoder (plus adapter) for high-fidelity input processing, discrete audio tokens remain essential for enabling the LLM to perform speech generation. To establish this output capability, we extend the LLM's vocabulary with discrete acoustic codes derived from StableToken (Song et al., 2025). This stage aligns text and audio representations via Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) tasks, effectively equipping the model with the interface to autoregressively predict audio segments for the speech decoder. During this process, only the embedding layer and the LLM head are trainable, while all other model parameters remain frozen.

Stage 2: Audio-LLM Adaptation. The second stage adapts the audio encoder to the pretrained LLM for cross-modal alignment. We train exclusively on UAS annotation data, with the LLM backbone and audio encoder frozen and only the projection layer updated. By introducing structured acoustic understanding at the outset, this stage prevents the model from developing ASR-centric representations that would later need to be "unlearned" to accommodate paralinguistic information.

Stage 3: Full Instruction Tuning. The third stage performs comprehensive instruction tuning across diverse audio understanding and generation tasks. We unfreeze all model parameters except the audio encoder and train on a mixture of: (1) **Foundational audio data**, following prior work (Xu et al., 2025a; Wu et al., 2025), including ASR and TTS tasks; (2) **UAS annotation** for generating complete UAS JSON from audio input; and (3) **UAS-QA** for answering questions about specific acoustic attributes across all UAS fields.

This diverse task mixture ensures that the model develops comprehensive audio intelligence spanning perception, reasoning, and generation. The inclusion of both UAS annotation and UAS-QA data—as validated by our ablation study—provides complementary supervision: UAS annotation teaches *what* to perceive, while UAS-QA teaches *how* to apply this knowledge.

Stage 4: GRPO. Following previous work (Wu et al., 2025), we utilize Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Li et al., 2025a) to further enhance the model's capabilities.

4 Experimental Setup

In this section, we detail the experimental setup, including the evaluation benchmarks, baseline models, and implementation details.

4.1 Evaluation Benchmarks

To comprehensively evaluate audio understanding and reasoning capabilities, we utilize benchmarks that collectively assess perception, reasoning, and generation abilities.

Audio Understanding. We evaluate audio understanding using three benchmarks. **MMSU** (Wang et al., 2025) is a comprehensive benchmark for understanding and reasoning in spoken language, comprising 5,000 audio-question-answer triplets across 47 tasks. Notably, MMSU is divided into *perception* and *reasoning* subsets, making it particularly relevant to our work as it directly measures the paralinguistic perception abilities that UAS targets. **MMAU** (Sakshi et al., 2024) evaluates multimodal audio understanding on tasks requiring expert-level knowledge and complex reasoning, comprising 10k audio clips with human-annotated QA pairs spanning speech, environmental sounds, and music. **MMAR** (Ma et al., 2025) evaluates deep reasoning capabilities of Audio-Language Models across multi-disciplinary tasks, containing 1,000 curated audio-question-answer triplets where

each item demands multi-step reasoning beyond surface-level understanding.

Audio Generation. We evaluate audio generation capabilities using **Seed-TTS** (Anastassiou et al., 2024) on both Chinese (Seed-zh) and English (Seed-en) test sets. This benchmark assesses the model's fundamental text-to-speech synthesis ability, ensuring that UAS training does not compromise basic generation quality.

4.2 Baselines

We compare against three state-of-the-art Audio-Language Models with 7B-scale language model backbones: Qwen2.5-Omni (Xu et al., 2025a), a unified multimodal model capable of perceiving and generating across text, images, audio, and video; Kimi-Audio (Ding et al., 2025), an audio-language model with strong audio understanding and generation capabilities; and Step-Audio2-mini (Wu et al., 2025), a compact yet powerful model designed for efficient audio understanding.

Results are cited from Xiaomi (2025) and we adopt the same evaluation framework.

4.3 Implementation Details

We build UAS-Audio upon the Qwen2.5-7B (Xu et al., 2025a) language model backbone with AuT (Audio Transformer) (Xu et al., 2025b) as our audio encoder, using a linear projection layer to align audio representations with the language model's embedding space. We use the AdamW optimizer with cosine learning rate scheduling and linear warmup across all stages. A comprehensive list of the specific datasets used for training is detailed in Appendix C. Detailed hyperparameter configurations for each training stage are provided in Appendix G.

While UAS serves as a dense supervision signal during training, we do not mandate the generation of full UAS JSONs during inference. Instead, we adopt a **task-specific prompting strategy** to align with standard evaluation protocols: we use standard transcription prompts for ASR, synthesis prompts for TTS, and discriminative prompts (e.g., multiple-choice) for understanding benchmarks like MMSU. This approach ensures our method remains compatible with existing metrics and incurs no additional latency overhead compared to task-specific baselines.

5 Results

5.1 Main Results

Table 1 presents comprehensive evaluation results across three benchmarks. Our UAS-Audio achieves the highest overall average of 65.2%, outperforming all baselines by a substantial margin.

Perception-Reasoning Trade-off. The most striking finding emerges from the MMSU benchmark, which explicitly decouples perception and reasoning capabilities. UAS-Audio achieves 55.7% on perception tasks—a remarkable **+10.9%** improvement over the best baseline (Kimi-Audio at 44.8%). Crucially, this perception gain does not come at the cost of reasoning: UAS-Audio maintains competitive reasoning accuracy (77.4%), only 0.2% below the top-performing Qwen2.5-Omni (77.6%). This result directly validates our hypothesis that ASR-centric training creates a perception bottleneck that UAS effectively addresses, while the structured schema format preserves the model's reasoning capabilities.

Cross-Domain Generalization. On the MMAR benchmark, which evaluates audio understanding across diverse domains, UAS-Audio achieves the highest overall score of 60.1%. Notably, UAS-Audio demonstrates consistent improvements in both speech-centric tasks (66.0%, +6.1% over Qwen2.5-Omni) and music understanding (45.2%, +4.4% over Qwen2.5-Omni), suggesting that our unified schema effectively captures domain-specific acoustic attributes. The improvement in speech tasks is particularly significant, as it indicates that enriching supervision with paralinguistic information enhances the model's ability to comprehend speaker characteristics and spoken content simultaneously.

Balanced Audio Understanding. On MMAU, while Step-Audio2 achieves the highest overall score (72.7%), UAS-Audio obtains competitive performance (69.4%) with notably balanced scores across all three categories (Speech: 67.0%, Sound: 70.0%, Music: 71.3%). This balance performance contrasts with baseline models that exhibit larger variance across domains, demonstrating that UAS provides a more uniform understanding of diverse audio types rather than excelling in specific categories at the expense of others.

Universality Across Architectures. We further validate the robustness of our approach by apply-

Model	MMSU			MMAR				MMAU				Avg.
	Perception	Reasoning	Overall	Speech	Sound	Music	Overall	Speech	Sound	Music	Overall	
Discrete Input Architecture												
GLM-4-Voice	11.04	16.16	13.30	34.35	29.70	19.90	29.60	35.44	27.63	27.84	30.30	24.4
UAS-Audio-D	31.32	48.55	39.66	44.56	40.00	36.89	43.30	46.25	59.16	62.57	56.00	44.2
Continuous Input Architecture												
Kimi-Audio	44.8	75.7	59.8	58.5	49.7	33.0	48.0	62.2	75.7	66.8	68.2	58.7
Qwen2.5-Omni	42.7	**77.6**	58.1	59.9	58.8	40.8	56.7	**70.6**	78.1	65.9	71.5	62.1
Step-Audio2	42.9	73.2	57.6	61.2	54.6	42.2	56.8	68.2	**79.3**	68.4	**72.7**	61.9
UAS-Audio	**55.7** (+10.9)	77.4	**66.2**	**66.0**	**58.8**	**45.2**	**60.1**	67.0	70.0	**71.3**	69.4	**65.2**

Table 1: Unified evaluation results on MMSU, MMAR, and MMAU benchmarks. The models are categorized by their audio input representation (Discrete vs. Continuous). **Bold** denotes the best result; underline denotes the second best. UAS-Audio-D denotes the discrete-input variant of UAS-Audio.

ing it to discrete audio representations (Top block of Table 1). Compared to the baseline GLM-4-Voice, our UAS-Audio-D achieves a transformative improvement, nearly doubling the overall average score (24.4% → 44.2%). This result confirms that the structured supervision of UAS is effective not only for high-fidelity continuous encodings but also for discrete token-based models.

5.2 Ablation

Figure 4 shows ablation results on MMSU. Both UAS annotation and UAS-QA contribute substantially to perception: removing UAS drops accuracy by 6.3%, removing UAS-QA by 9.6%, and removing both by 15.0%. The larger impact of UAS-QA suggests that explicit question-answering training is more critical for translating acoustic knowledge into task performance.

Crucially, reasoning accuracy remains stable across all configurations, confirming that our perception enhancements do not compromise reasoning capabilities. This validates our hypothesis that perception and reasoning operate as independent factors in audio understanding, and that current Audio LLMs suffer primarily from a perception bottleneck rather than reasoning limitations.

Due to space limitations, further analyses in the appendices ablate the impact of GRPO (Appendix E) and demonstrate that our structured schema significantly outperforms unstructured captions (Appendix F).

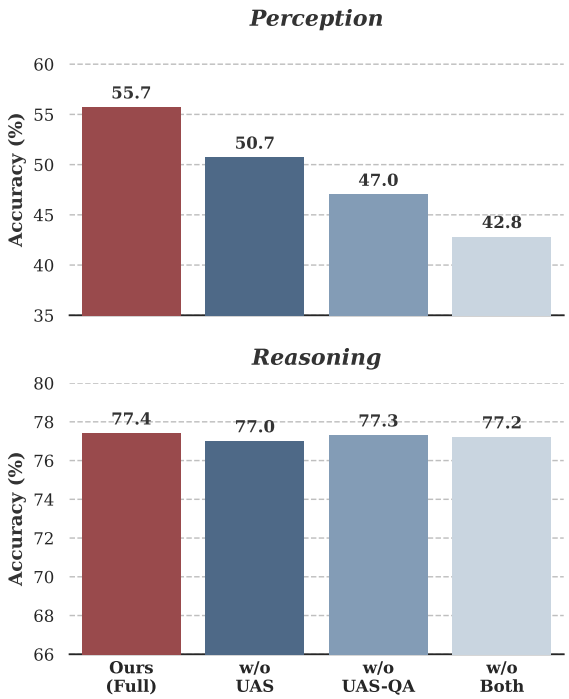


Figure 4: **Impact of individual components**. We observe a consistent performance drop when removing the UAS module or the UAS-QA strategy from the full model.

5.3 Speech Generation Ability

To verify that perception-focused training does not compromise generation capabilities, we evaluate TTS performance on Seed-TTS benchmarks (Table 2). UAS-Audio achieves the best average WER score of 1.6, outperforming Qwen2.5-Omni (1.9) and Step-Audio2-mini (2.7). Notably, UAS-Audio matches or exceeds baselines on both Chinese and English synthesis, demonstrating that our UAS training not only preserves but enhances speech generation quality. We attribute this to the enriched acoustic representations learned through structured paralinguistic supervision, which naturally transfer

Model	Seed-Zh	Seed-EN	Avg
Baichuan-Audio-Instruct	2.9	4.7	3.8
Qwen2.5-Omni	1.4	2.3	1.9
Step-Audio2-mini	2.1	3.2	2.7
UAS-Audio	**1.4**	**1.7**	**1.6**

Table 2: Evaluation results on TTS benchmarks. **Avg** reports the average score of Seed-Zh and Seed-EN. **Bold** denotes the best result.

Prompt Strategy	Output	LS-Clean	AISHELL
Targeted ASR	Raw Text	2.2	2.3
Holistic UAS	**JSON**	**2.3**	**2.4**
Performance Deviation		**≈ 0.1**	**≈ 0.1**

Table 3: **Robustness of ASR Capability under Multi-Task Generation.** Comparing standard ASR generation versus extracting transcription from the unified JSON output (Holistic). The model maintains high recognition accuracy even when handling complex structured predictions.

to benefit generation tasks. This confirms that UAS-Audio achieves unified audio intelligence without understanding-generation trade-offs.

5.4 Flexibility and Robustness of Structured Generation.

While UAS-Audio supports standard, low-latency ASR generation (Targeted Mode), it uniquely possesses the capability to output holistic UAS JSONs—providing rich acoustic context alongside the transcript. A natural question is whether embedding transcription within such a complex structure imposes a "tax" on recognition accuracy. To investigate this, we treat the holistic UAS generation as a stress test and compare its transcription field against standard ASR prompting.

As shown in Table 3, the transcription quality remains virtually identical across both settings. On the LibriSpeech test-clean set(Panayotov et al., 2015), the WER difference is merely 0.1 (2.2 vs. 2.3), and similarly on AIShell (Bu et al., 2017), the difference is also 0.1 (2.3 vs. 2.4). This negligible deviation demonstrates that our model maintains high recognition accuracy even when simultaneously predicting paralinguistic attributes, speaker characteristics, and non-linguistic events within a structured JSON format.

6 Related Work

The evolution of LLMs has driven the transition of spoken dialogue models from traditional cascaded pipelines to end-to-end Audio-Language Models (AudioLLMs) (Lakhotia et al., 2021; Rubenstein et al., 2023; Zhang et al., 2023; Gong et al., 2024; Tang et al., 2023; Hu et al., 2024; Fang et al., 2024; Défossez et al., 2024; Li et al., 2025b; Wang et al., 2024; Bai et al., 2024; Goel et al., 2025; Wijngaard et al., 2025). These models generally adopt one of two audio representation strategies: continuous representations (Huang et al., 2025) or discrete representations (Van Den Oord et al., 2017; Zeng et al., 2024; Song et al., 2025). Despite significant progress in speech recognition and dialogue, most existing AudioLLMs primarily focus on *what* is spoken while underutilizing paralinguistic information about *how* it is spoken.

Recent works have explored richer audio representations. MiDashengLM (Dinkel et al., 2025) proposes using general audio captions to fuse speech transcripts with acoustic descriptions into a unified textual output. However, this approach relies on *unstructured* natural language, which inherently entangles paralinguistic nuances with linguistic content rather than isolating them. In contrast to such implicit supervision, our work advocates for an explicitly structured schema to ensure precise disentanglement of perception and reasoning. SenseVoice (An et al., 2024) implements rich transcription through interleaved tags, inserting special tokens (e.g., <laughter>, <happy>) directly into the linear ASR text stream. While useful for sparse events, this "flat" tagging structure generalizes poorly to dense, continuous attributes such as prosody, timbre, and emotion. Additionally, it treats these representations as specialized outputs rather than a general-purpose supervision methodology for AudioLLM training.

7 Conclusion

We propose the Unified Audio Schema (UAS) to address the deficiency of fine-grained perception in current AudioLLMs. By restructuring supervision across linguistic, paralinguistic, and environmental dimensions, UAS-Audio achieves a 10.9% improvement in perception accuracy on the MMSU benchmark. Importantly, this gain is achieved without compromising reasoning or transcription capabilities. Our findings suggest that structural richness in supervision is a critical factor for evolving AudioLLMs beyond simple scaling.

Limitations

Despite the promising results of UAS-Audio in bridging the perception-reasoning gap, we acknowledge certain limitations in our current study that point to directions for future research:

- **Linguistic Diversity:** Our current experimental validation primarily focuses on high-resource languages (English and Chinese) due to the composition of mainstream benchmarks like MMSU and Seed-TTS. While the UAS framework is language-agnostic by design, its efficacy on low-resource or code-switching scenarios remains to be verified in future work.
- **Complex Overlapping Speech:** While UAS effectively handles background environmental events, our current schema focuses on the primary speaker for paralinguistic analysis. Scenarios involving highly overlapping speech (e.g., the cocktail party problem) with multiple active speakers requiring simultaneous paralinguistic disentanglement are not fully covered in this iteration and warrant further investigation.

References

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Haoneng Luo, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, and 14 others. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *Preprint*, arXiv:2407.04051.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, and 1 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Jisheng Bai, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, Mark D. Plumbley, Woon-Seng Gan, and Jianfeng Chen. 2024. Audiosetcaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models. *Preprint*, arXiv:2411.18953.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. 2021. Hi-fi multi-speaker english tts dataset. *arXiv preprint arXiv:2104.01497*.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pages 1–5. IEEE.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. 2025. Midashenglm: Efficient audio understanding with general audio captions. *arXiv preprint arXiv:2508.03983*.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA.

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

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. 2024. Listen, think, and understand. *Preprint*, arXiv:2305.10790.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, and 1 others. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, and 1 others. 2024. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, and 1 others. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Yejin Kwon, Taewoo Kang, Hyunsoo Yoon, and Changouk Kim. 2025. M3-slu: Evaluating speaker-attributed reasoning in multimodal large language models. *Preprint*, arXiv:2510.19358.

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and 1 others. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

John Laver. 1994. *Principles of Phonetics*. Cambridge University Press.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. 2025a. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, and 1 others. 2025b. Baichuan-audio: A unified framework for end-to-end speech interaction. *arXiv preprint arXiv:2502.17239*.

Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. 2023. Yodas: Youtube-oriented dataset for audio and speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *ICLR*. OpenReview.net.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, and 1 others. 2025. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *Proc. NeurIPS*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, and 1 others. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Yuhan Song, Linhao Zhang, Chuhan Wu, Aiwei Liu, Wei Jia, Houfeng Wang, and Xiao Zhou. 2025. Stabletoken: A noise-robust semantic speech tokenizer for resilient speechllms. *Preprint*, arXiv:2509.22220.

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*.

Aaron Van Den Oord, Oriol Vinyals, and 1 others. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. 2017. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.

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

Xiong Wang, Yangze Li, Chaoyou Fu, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2024. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm. *arXiv preprint arXiv:2411.00774*.

Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. 2025. Audsemthinker: Enhancing audio-language models through reasoning over semantics of sound. *Preprint*, arXiv:2505.14142.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, and 1 others. 2025. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

LLM-Core-Team Xiaomi. 2025. Mimo-audio: Audio language models are few-shot learners.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. 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. *Preprint*, arXiv:2402.07729.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, and 1 others. 2022. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186. IEEE.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *Preprint*, arXiv:2305.11000.

A Human Evaluation of UAS Data Quality

To rigorously assess the reliability of the automated UAS generation pipeline, we conducted a manual quality audit on a randomly sampled subset of the generated dataset.

Data Sampling and Participants We randomly sampled $N = 200$ audio segments, ensuring a representative distribution across diverse audio types, including speech, music, and environmental sounds. The evaluation involved three volunteer human annotators with expertise in audio analysis. All participants provided informed consent prior to the task. They were informed that the task involved listening to standard audio samples and that they could withdraw at any time. No monetary compensation was provided.

Annotation Interface and Instructions To facilitate the evaluation process, we developed a web-based annotation interface, as shown in Figure 5. The interface presents the audio player at the top, followed by a series of predicted attributes (e.g., Age, Gender, Emotion) derived from the synthesized UAS JSON.

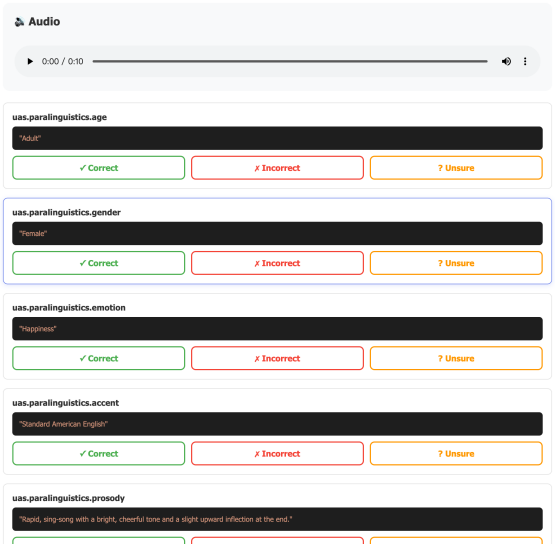


Figure 5: The web-based human evaluation interface. Annotators listen to the audio and verify if the predicted JSON values (black boxes) match the acoustic reality by selecting "Correct", "Incorrect", or "Unsure".

The instructions provided to the participants are as follows:

> *Please listen to the audio clip provided in the player. Below the player, you will see a list of attributes extracted from the system. Please determine if the label accurately describes the audio content by selecting one of the options: Correct, Incorrect, or Unsure.*

Evaluation Metric For each sample, three annotators independently verified the alignment between the audio signal and the synthesized UAS JSON fields. An attribute was marked as "correct" only if the annotator consensus (majority vote) confirmed it accurately reflected the acoustic reality. We report the mean accuracy across the nine specific fields within the Paralinguistics and Non-linguistic Events domains.

Quantitative Results As shown in Table 4, the UAS pipeline achieves high precision across most fields, with most paralinguistic and environmental attributes exceeding 95% accuracy.

UAS Domain	Field	Accuracy (%)	95% CI
Paralinguistics	Age	98.5	[95.7, 99.5]
	Gender	97.0	[93.6, 98.6]
	Emotion	89.5	[84.5, 93.1]
	Accent	96.0	[92.3, 98.0]
	Prosody	95.5	[91.7, 97.6]
	Timbre	95.0	[91.0, 97.3]
Non-linguistic Events	Description	96.5	[92.9, 98.3]
	Discrete Events	90.0	[85.1, 93.4]
	Continuous Events	96.0	[92.3, 98.0]

Table 4: Human evaluation accuracy across UAS fields ($N = 200$). 95% Confidence Intervals (CI) are calculated using the Wilson score interval method.

Analysis The evaluation reveals that stable biological and environmental traits (e.g., *Age*, *Gender*, and *Continuous Events*) exhibit a high degree of reliability, with lower bounds of the 95% confidence intervals consistently remaining above 92%. This suggests that the pipeline effectively captures these relatively objective acoustic properties. The relative performance decrease in *Emotion* (89.5%, 95% CI: [84.5, 93.1]) and *Discrete Events* (90.0%, 95% CI: [85.1, 93.4]) reflects the inherent subjectivity of emotional state perception and the temporal sparsity of short-duration sound events. However, even accounting for statistical uncertainty, the accuracy for these challenging fields remains robustly above 84%, demonstrating that the UAS pipeline provides a high-fidelity representation across all acoustic dimensions.

B Discrete Architecture Training

For the discrete architecture, we initialize the model with Qwen2.5-3B and adopt the same audio tokenizer as GLM-4-Voice (Zeng et al., 2024).

Compared to the continuous architecture, the training pipeline is simplified in two aspects.

First, we omit the adapter alignment stage, since discrete audio tokens are directly embedded into the LLM vocabulary and therefore do not require an additional projection or alignment module.

Second, we do not employ GRPO-style reinforcement learning. This choice follows the training design of GLM-4-Voice, where discrete audio representations are learned purely through supervised objectives without reinforcement learning.

The model is trained on UAS annotation tasks to establish structured acoustic understanding, followed by fine-tuning on UAS-QA tasks to enhance its ability to answer questions about specific acoustic attributes. The learning rate is scheduled with cosine decay, starting from 1×10^{-4} and 5×10^{-5} respectively.

C Training Datasets for UAS-Audio

We train UAS-Audio on hundreds of thousands of hours of both open-source data and in-house data. All open-source datasets used in this work are listed in Table 5.

Dataset	Duration (#hours)
LibriSpeech (Panayotov et al., 2015)	960
Multilingual LibriSpeech (Pratap et al., 2020)	27,322
GigaSpeech (Chen et al., 2021)	10,000
Yodas (Li et al., 2023)	29,155
Hi-Fi TTS (Bakhturina et al., 2021)	292
VCTK (Veaux et al., 2017)	44
LibriTTS (Zen et al., 2019)	586
AISHELL-1 (Bu et al., 2017)	150
WenetSpeech (Zhang et al., 2022)	10,005
Common Voice (Ardila et al., 2019)	2,133
Emilia (He et al., 2024)	96,750
AudioSet (Gemmeke et al., 2017)	4,922

Table 5: Summary of datasets used for training UAS-Audio

D Inference Efficiency and Flexible Generation Strategy

A potential concern regarding the adoption of the Unified Audio Schema (UAS) is the inference latency and token overhead associated with generating verbose JSON structures. While the full UAS format (containing keys for transcription, paralinguistics, and events) involves a larger number of output tokens compared to simple tagging, it is crucial to distinguish between UAS as a *supervision objective* during training and UAS as an *inference format* during deployment.

Decoupling Supervision from Generation. The primary goal of training with UAS is to force the model to internally disentangle and encode fine-grained acoustic information that is typically suppressed by ASR-centric objectives. Once the model is aligned via this structured supervision (specifically after Stage 3: Full Instruction Tuning), the rich acoustic representations are embedded within the model's parameters. Consequently, during inference, the generation format is entirely flexible and controlled by the user prompt.

Targeted Perception via Prompting. As demonstrated by the inclusion of the UAS-QA dataset in our training pipeline, UAS-Audio is capable of following diverse instructions. For latency-sensitive applications, users need not generate the full holistic JSON. Instead, they can prompt the model to extract specific attributes directly.

For instance, to retrieve emotion and gender, a user can prompt: *"Identify the speaker's emotion and gender in a concise format."* The model, having learned the underlying concepts through the UAS schema, can output a short response (e.g., *"Neutral, Male"*) without the overhead of the full JSON syntax. This "Targeted Mode" drastically reduces token consumption to be comparable with standard classification heads, while still benefiting from the superior perception accuracy gained from the UAS training paradigm.

In summary, the structured JSON serves as a high-density information scaffold during learning, but the model retains the flexibility to operate in a low-latency, token-efficient manner during inference.

E Impact of GRPO

To explicitly quantify the contribution of the reinforcement learning stage, we evaluate a variant of UAS-Audio trained without Stage 4 (GRPO). As presented in Table 6, removing the GRPO stage results in a marginal performance drop of 0.9% in perception (55.7% $\rightarrow$ 54.8%) and 1.4% in reasoning (77.4% $\rightarrow$ 76.0%).While these results confirm that GRPO serves as an effective strategy for improving model performance, the ablation highlights a critical finding: even without GRPO, UAS-Audio

achieves a perception accuracy of 54.8%, which still surpasses the strongest baseline (Kimi-Audio, 44.8%) by a significant margin of 10.0%. This empirical evidence demonstrates that the substantial performance leap (+11%) reported in our main results is primarily driven by the structured supervision of the Unified Audio Schema (UAS), rather than the optimization techniques in the final training stage.

Model Settings	Perception	Reasoning	Average
UAS-Audio (Full)	**55.7**	**77.4**	**66.2**
− w/o GRPO (Stage 4)	54.8	76.0	65.2
Best Baseline (Kimi-Audio)	*44.8*	*75.7*	*62.2*

Table 6: Ablation study on the impact of GRPO training (Stage 4) on the MMSU benchmark. The results show that while GRPO provides further optimization, the majority of the performance gain stems from the UAS supervision.

F Impact of Structured Format

To validate the effectiveness of our schema design, we conducted a controlled comparison between **UAS Supervision** (structured JSON) and a **Caption Supervision** baseline (unstructured natural language). Notably, this Caption Supervision setting conceptually emulates the *general audio captioning* paradigm adopted by recent works such as MiDashengLM (Dinkel et al., 2025). We exclude the final reinforcement learning stage (Stage 4) in both settings to strictly isolate the impact of the supervision format while controlling for model architecture and training data.

Target Format	Perception	Reasoning	Average
Unstructured Caption	48.4	75.5	61.5
Structured UAS	**54.8**	**76.0**	**65.2**

Table 7: Impact of supervision format on MMSU. The "Unstructured Caption" setting serves as a proxy for caption-based approaches (e.g., MiDashengLM). Both settings use the exact **same synthetic data source**. All models are trained without GRPO.

As shown in Table 7, structured UAS yields a significant 6.4% perception gain over the unstructured caption baseline. confirming that the schema format itself lowers learning difficulty given identical data. Specifically, UAS enforces **explicit disentanglement** by assigning orthogonal slots to prevent semantic interference, offers **syntactic invariance** by providing a low-entropy target devoid of linguistic variability, and ensures **forced completeness** by mandating dense predictions for all acoustic fields, thereby capturing subtle details often omitted in fluent captions.

G Training Hyperparameters

Table 8 summarizes the hyperparameter configurations for each training stage.

H Prompts Used in Experiments

In this section, we provide the detailed prompts used in our pipeline for reproducibility.

H.1 Audio Caption to UAS Format Conversion

We utilized the Qwen3-30B-A3B-Instruct model to convert raw audio captions into the Unified Audio Description (UAS) format. The prompt used for this transformation is shown in Figure 6.

H.2 QA Pair Generation from UAS

To generate Question-Answer (QA) pairs based on the audio UAS format descriptions, we employed the Qwen3-235B-A22B-Instruct model. The specific prompt guiding this generation process is detailed in Figure 7.

I LLM Usage Statement

In accordance with the conference policies on Large Language Model (LLM) usage, we hereby disclose the following: After completing the initial draft of this paper, we utilized LLMs to enhance grammar and polish the writing of this manuscript. No new research ideas, experimental designs, or scientific content were generated by LLMs.

This statement is provided to ensure transparency and compliance with the conference's policies on LLM usage.

Hyperparameter	Stage 1	Stage 2	Stage 3	Stage 4
Peak Learning Rate	5e-4	2e-4	1e-4	5e-6
Warmup Iterations	500	1,000	1,000	200
LR Schedule	Cosine with Linear Warmup			
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$)			
Weight Decay	0.1			
Gradient Clipping	1.0			
Trainable Parameters	Projector	Projector	All (excl. Encoder)	All (excl. Encoder)

Table 8: Hyperparameter configurations across training stages.

Prompt for Caption-to-UAS Conversion

```
Given a detailed description of an audio sample, output a JSON object containing the
    following audio features:

- **transcription**: If human speech is present, provide an accurate transcription of the
    spoken content in the original language. If there is no human voice, set this field to
    null.
- **paralinguistics**: If human voice is present, provide the following fields:
  - `age`: One of `Child`, `Adult`, or `Elderly`.
  - `gender`: Specify as `Male` or `Female`.
  - `emotion`: This field MUST use ONE of the following seven specific categories: `Anger`,
    `Disgust`, `Sadness`, `Happiness`, `Neutral`, `Surprise`, `Fear`. Only these values are
    allowed.
  - `accent`: Describe the accent or variety of language used (e.g., `Standard Mandarin
    Chinese`, `American English`, etc.).
  - `prosody`: Summarize prosodic features, which refer to the patterns of rhythm, pitch,
    pace, emphasis, and intonation in speech (i.e., how something is said).
  - `timbre`: Briefly describe the timbre of the voice. **Timbre** refers to the unique tonal
    quality or color of a sound that distinguishes one voice or instrument from another,
    independent of pitch and loudness. For example, descriptors may include "nasal,"
    "breathy," "warm," "bright," "harsh," or "gentle."

  **Note:** Timbre is *not* the same as prosody; prosody relates to temporal and pitch-based
    features, while timbre describes the characteristic sound qualities.

  If there is no human voice, set all fields in the `paralinguistics` object to null.
- **nonLinguisticEvents**:
  - `description`: A summary sentence describing general non-speech audio characteristics or
    context.
  - `discreteEvents`: A list of discrete (one-shot or instantaneous) non-linguistic events
    (such as a car horn, a door slam). Each item must contain a unique `label` and a brief
    `characteristic` describing its intensity, duration, or other relevant attribute. (e.g.,
    `label`: `"Car horn"`, `characteristic`: `"Short, loud"`). Event labels must not repeat.
  - `continuousEvents`: A list of continuous or background non-linguistic events (such as
    engine noise, wind, music), again with a unique `label` and a brief `characteristic`
    descriptor.

Always follow these rules:
- If the audio contains **no human voice**, set `transcription` and all fields inside
    `paralinguistics` to null.
- For `emotion`, ONLY USE ONE OF THESE: `Anger`, `Disgust`, `Sadness`, `Happiness`,
    `Neutral`, `Surprise`, `Fear`.
- Ensure that all event `labels` are unique and clearly indicate what type of sound or event
    they refer to.

Respond ONLY with a JSON object as output (do not include any preamble, explanation, or extra
    formatting), with all required fields. Use the formats and categories exactly as
    described above.
```

Figure 6: Prompt for using the Qwen3-30B-A3B-Instruct model to perform Caption-to-UAS Conversion

Prompt for QA Generation

```
**Instructions:**
You are given a structured audio description in UAS (Unified Audio Schema) JSON format.
    Please generate a relevant question in the form of a **Multiple Choice** question, along
    with the corresponding answer, based on the specific fields provided in the JSON (such as
    transcription, paralinguistics, or non-linguistic events).

**Requirements:**
- Provide 3-4 answer options. Each option must include both the letter and the content (e.g.,
    "A. male", "B. female").
- The question can pertain to specific attributes found in the UAS structure, such as:
  - The speaker's gender, age, emotion, accent, prosody, and timbre (from `paralinguistics`).
  - Specific sounds or events (from `discreteEvents` or `continuousEvents`).
  - The content of speech (from `transcription`).
- The question text must not directly reveal or hint at the answer; answering must require
    information from the audio, and not be possible by simply reading the question.
- Do not include phrases like "according to the JSON" or "in the paralinguistics field".
- The correct answer must be option ${correct_option}.

**Input Format:**
A JSON object containing `transcription`, `paralinguistics`, and `nonLinguisticEvents`.

**Output Format:**
Present your output in the following JSON format:
```json
[
    {"role": "user", "content": [{"type": "text", "text": "question_text"}]},
    {"role": "assistant", "content": "answer_text"}
]
```

**Now, generate a question and its answer for the following UAS input using the above
    guidelines:**

${uas}
```

Figure 7: Prompt for using the Qwen3-235B-A22B-Instruct model to perform QA Generation from UAS input